ANÁLISIS Y COMPARACIÓN DE MODELOS DE APRENDIZAJE AUTOMÁTICO, ESTADÍSTICO Y MATEMÁTICO PARA LA PREDICCIÓN DE BROTES EN SALUD PÚBLICA.

DEISY FORERO BENAVIDES
JEISSON RODRÍGUEZ RODRÍGUEZ
ZUJEL ENRIQUE ROMERO PÉREZ

Proyecto aplicado para optar al título de Magíster en Ciencia de Datos

Directora
Dra. DELIA ORTEGA LENIS

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, ENERO 19 DE 2026

TABLA DE CONTENIDO

INTRODUCCIÓN 8

1. OBJETIVOS DEL PROYECTO 10

1.1. OBJETIVO GENERAL 10

1.2. OBJETIVOS ESPECÍFICOS 10

2. MARCO TEÓRICO Y ANTECEDENTES 11

2.1. MARCO TEÓRICO 11

2.1.1. Modelos Matemáticos 12

2.1.2. Modelos Estadísticos 15

2.1.3. Modelos de Aprendizaje Automático (Machine Learning) 19

2.2. ANTECEDENTES 21

3. PREPARACIÓN, LIMPIEZA Y ORGANIZACIÓN DE DATOS ABIERTOS DE COVID19 25

3.1. OBTENCIÓN DEL CONJUNTO DE DATOS 25

3.2. DESCRIPCIÓN GENERAL DEL CONJUNTO DE DATOS 25

3.3. DETECCIÓN DE PROBLEMAS Y ACCIONES CORRECTIVAS 27

3.3.1. Nombres de columnas 27

3.3.2. Eliminación de columnas 27

3.3.3. Tratamiento de valores nulos 28

3.3.4. Estandarización de variables categóricas 29

3.3.5. Validación de fechas 29

3.3.6. Duplicados 29

3.3.7. Agrupación de serie de tiempo 29

3.3.8. Transformaciones. 31

3.3.9. Definición de tramos u olas 31

3.3.10. Análisis de descomposición por tramos. 32

3.4. RESULTADO DE LA LIMPIEZA 35

4. IDENTIFICACIÓN Y SELECCIÓN DE LAS VARIABLES CRÍTICAS 36

4.1. SELECCIÓN DE VARIABLES 36

4.2. ESTRATEGIAS DE SELECCIÓN DE VARIABLES 36

4.3. ANÁLISIS DE CORRELACIÓN DE VARIABLES EXOGENAS 36

4.4. INGENIERÍA DE CARACTERÍSTICAS 37
5. ENTRENAMIENTO Y ESTIMACIÓN DE LOS MODELOS 39
5.1. MODELO MATEMÁTICO 39
5.1.1. MODELO PRIMER TRAMO 42
5.1.2. MODELO SEGUNDO TRAMO 44
5.1.3. MODELO TERCER TRAMO 45
5.1.4. MODELO CUARTO TRAMO 47
5.2. MODELOS ESTADÍSTICOS. 50
5.2.1. MODELOS PRIMER TRAMO 50
5.2.2. MODELOS SEGUNDO TRAMO 57
5.2.3. MODELOS TERCER TRAMO 64
5.2.4. MODELOS CUARTO TRAMO 70
5.3. MODELOS DE APRENDIZAJE AUTOMATIVO 73
5.3.1. MODELOS PRIMER TRAMO. 75
5.3.2. MODELOS SEGUNDO TRAMO 77
5.3.3. MODELOS TERCER TRAMO. 79
5.3.4. MODELOS CUARTO TRAMO. 81
5.3.5. TRAMOS ACUMULADOS 84
6. COMPARACIÓN DE LOS MODELOS 88
6.1. PRIMER TRAMO: FASE INICIAL DE ALTA INCERTIDUMBRE 88
6.2. SEGUNDO TRAMO: CONSOLIDACIÓN DE PATRONES TEMPORALES 89
6.3. TERCER TRAMO: FASE DE MÁXIMA INTENSIDAD EPIDÉMICA 90
6.4. CUARTO TRAMO: FASE DE DESCENSO Y BAJA INCIDENCIA 91
7. CONCLUSIONES Y LIMITACIONES 93
7.1. CONCLUSIONES 93
7.2. LIMITACIONES 94
REFERENCIAS BIBLIOGRÁFICAS 96

INDICE DE FIGURAS

figura 1. Evolución temporal de infectados con tramos destacados 31
figura 2. Descomposición estacional del primer tramo 33
figura 3. Descomposición estacional del segundo tramo 33
figura 4. Descomposición estacional tercer tramo 34
figura 5. Descomposición estacional cuarto tramo 34
figura 6. Matriz de correlación de las variables 36
figura 7. Gráfica SEIR continuo 41
figura 8. Gráfica primer tramo 43
figura 9. Predicción primer tramo 43
figura 10. Gráfica segundo tramo 44
figura 11. Predicción segundo tramo 45
figura 12. Gráfica tercer tramo 46
figura 13. Predicción tercer tramo 46
figura 14. Gráfica cuarto tramo 47
figura 15. Predicción cuarto tramo 48
figura 16. Predicción primer tramo 51
figura 17. Diagnóstico de Residuos del Modelo SARIMA+GARCH 55
figura 18. Pronóstico modelo Arima segundo tramo 59
figura 19. Pronóstico modelo Sarima segundo tramo 62
figura 20. Pronóstico modelo Sarimax segundo tramo 63
figura 21. Ajuste modelo Arima tercer tramo 65
figura 22. Pronóstico modelo Sarima tercer tramo 67
figura 23. Pronóstico modelo Sarimax tercer tramo 69
figura 24. Sarimax tercer tramo 70
figura 25. Arima cuarto tramo 71
figura 26. Sarima cuarto tramo 72
figura 27. Sarimax cuarto tramo 73
figura 28. Serie completa primer tramo vs predicción modelo XGBoost 76
figura 29. Rango de las variables utilizadas primer tramo 76
figura 30. Serie completa primer tramo vs predicción LSTM 77
figura 31. Serie completa segundo tramo vs Predicción con modelo XGBoots 78
figura 32. Rango de las variables utilizadas segundo tramo 78
figura 33. Seria completa segundo tramo vs predicción LSTM 79
figura 34. Serie completa tercer tramo vs Predicción con modelo XGBoots 80
figura 35. Rango de las variables utilizadas tercer tramo 80
figura 36. Serie completa tercer tramo vs predicción LSTM 81
figura 37. Serie completa cuarto tramo vs Predicción con modelo XGBoots 82
figura 38. Rango de las variables utilizadas cuarto tramo 82
figura 39. Seria completa cuarto tramo vs predicción LSTM 83
figura 40. Análisis de tramos acumulados de los modelos XGBoost. 85
figura 41. Análisis de tramos acumulados de los modelos LSTM. 86

figura 42. Comparación de métricas por modelo primer tramo 89
figura 43. Comparación de métricas por modelo segundo tramo 90
figura 44. Comparación de métricas por modelo tercer tramo 91
Figura 45. Comparación de métricas por modelo cuarto tramo 92

INDICE DE TABLAS

Tabla 1. Identificación de las variables 26
Tabla 2. Registros faltantes en la Base de datos COVID19 28
Tabla 3. Clasificación de variables 31
Tabla 4. Fechas de tramos temporales para COVID19 32
Tabla 5. Prueba de Dickey-Fuller Aumentada (ADF) y análisis de estacionariedad por tramos 32
Tabla 6. Características de las variables creadas. 37
Tabla 7. Parámetros iniciales modelo continuo 41
Tabla 8. Métricas modelo continuo 41
Tabla 9. Parámetros primer tramo 42
Tabla 10. Métricas primer tramo 43
Tabla 11. Métricas de predicción primer tramo 44
Tabla 12. Parámetros segundo tramo 44
Tabla 13. Métricas segundo tramo 44
Tabla 14. Métricas de predicción segundo tramo 45
Tabla 15. Parámetros tercer tramo 45
Tabla 16. Métricas tercer tramo 46
Tabla 17. Métricas de predicción tercer tramo 47
Tabla 18. Parámetros cuarto tramo 47
Tabla 19. Métricas cuarto tramo 47
Tabla 20. Métricas de predicción cuarto tramo 48
Tabla 21. Resumen de parámetros y métricas 48
Tabla 22. Métricas MAPE y MASE 49
Tabla 23. Métricas del modelo 53
Tabla 24. Parámetros del modelo GARCH(1,1) 54
Tabla 25. Pruebas de Diagnóstico de Residuos 55
Tabla 26. Métricas de Precisión del Modelo Final 55
Tabla 27. Combinaciones modelo sarima 56
Tabla 28. Métricas modelo Sarimax 56
Tabla 29. Métricas Arima 58
Tabla 30. Parámetros estimados 58
Tabla 31. Métricas del modelo 59
Tabla 32. Comportamiento de los ARIMA 59
Tabla 33. ACF y PACF del modelo Sarima 61
Tabla 34. Métricas del modelo SARIMAX 62
Tabla 35. Comparación Arima y Sarima tercer tramo 66
Tabla 36. Parámetros Sarima tercer tramo 67
Tabla 37. Comparación Arima y Sarima tercer tramo 67
Tabla 38. Interpretación de las variables 68
Tabla 39. Métricas modelo cuarto tramo 70
Tabla 40. Métricas modelo Sarima cuarto tramo 71
Tabla 41. Métricas modelos cuarto tramo 72

Tabla 42. Boxcox primer tramo 75
Tabla 43. Minmax primer tramo 76
Tabla 44. Métricas de evaluación primer tramo 77
Tabla 45. Boxcox segundo tramo 77
Tabla 46. Minmax segundo tramo 78
Tabla 47. Métricas de evaluación segundo tramo 79
Tabla 48. Minmax tercer tramo 79
Tabla 49. Tercer tramo 81
Tabla 50. Métricas de evaluación tercer tramo 81
Tabla 51. Minmax cuarto tramo 82
Tabla 52. Métricas de los modelos XGBoost. 83
Tabla 53. Métricas de evaluación cuarto tramo 83
Tabla 54. Hiperparámetros óptimos del modelo XGBoost con tramos acumulados 85
Tabla 55. Métricas del modelo XGBoost con tramos acumulados. 86
Tabla 56. Hiperparámetros óptimos del modelo LSTM con tramos acumulados. 87
Tabla 57. Métricas del modelo LSTM con tramos acumulados. 87
Tabla 58. Primer tramo 88
Tabla 59. Segundo tramo 89
Tabla 60. Tercer tramo 90
Tabla 61. Cuarto tramo 91

INTRODUCCIÓN

Los brotes de enfermedades infecciosas han representado históricamente uno de los desafíos más complejos y persistentes para la salud pública global, debido a su capacidad de propagación rápida, su impacto sobre la mortalidad y la presión que ejercen sobre los sistemas sanitarios (1), (2). La pandemia de COVID19, iniciada a finales de 2019, evidenció de manera contundente las vulnerabilidades estructurales de los sistemas de salud a nivel mundial, destacando la necesidad de respuestas rápidas y efectivas ante emergencias sanitarias de gran magnitud (3), (4). De forma paralela, los brotes recurrentes como el dengue, la gripe estacional y el cólera continúan afectando la estabilidad sanitaria, especialmente en regiones con recursos limitados, donde factores sociales, ambientales y económicos influyen de manera significativa en su propagación (5) , (6). Este panorama ha reafirmado la urgencia de contar con herramientas predictivas capaces de anticipar la aparición y evolución de enfermedades infecciosas (7).

En este contexto, la predicción de brotes se ha consolidado como una prioridad estratégica en salud pública. La identificación temprana de tendencias emergentes, zonas de riesgo y posibles picos de contagio no solo permite mitigar el impacto sanitario, sino también optimizar el uso de recursos y reducir pérdidas económicas y sociales (2), (8). Para ello, la ciencia de datos se ha posicionado como un recurso clave, al permitir el análisis de grandes volúmenes de información histórica y actual con el fin de identificar patrones, comportamientos no lineales y dinámicas complejas que faciliten una respuesta sanitaria proactiva y basada en evidencia (9), (10) (11).

Para el desarrollo del estudio se determinó acotar el alcance geográfico exclusivamente a la ciudad de Bogotá, con el propósito de lograr una mayor profundidad en el tratamiento de los datos disponibles y un análisis más preciso del comportamiento epidemiológico en un contexto urbano altamente densificado. Este ajuste permitió delimitar mejor el contexto analítico, aunque también planteó nuevos desafíos metodológicos. Entre ellos, se encuentra el proceso de identificación, depuración y validación de las variables más relevantes para la construcción de los modelos predictivos, proceso que requiere contrastar diversas fuentes de información y evaluar la pertinencia de los datos según los enfoques seleccionados (12), (13).

A pesar de ello, se avanzó en el diseño preliminar de modelos predictivos para brotes de enfermedades infecciosas, comparando tres enfoques metodológicos ampliamente utilizados en la literatura: el modelado matemático, el modelado estadístico y el aprendizaje automático. El enfoque matemático ha demostrado ser útil para representar la dinámica de transmisión mediante modelos compartimentales como SEIR (14), (15), mientras que los modelos estadísticos permiten capturar tendencias temporales y patrones históricos en los datos epidemiológicos (12), (13). Por su parte, las técnicas de aprendizaje automático ofrecen ventajas en el procesamiento de grandes volúmenes de datos y en la identificación de relaciones complejas no explícitas entre variables (9), (16). Este análisis comparativo permitió identificar fortalezas, limitaciones y condiciones mínimas necesarias para su implementación en contextos reales de salud pública local.

En un entorno globalizado e influido por el cambio climático, donde las amenazas epidemiológicas se

intensifican y diversifican, contar con modelos predictivos robustos se ha vuelto esencial para fortalecer la capacidad de respuesta sanitaria (1).

En consecuencia, se espera que sus aportes contribuyan al conocimiento académico y sirvan como insumo práctico para el fortalecimiento de la gestión sanitaria, no solo en la ciudad de Bogotá, sino también replicable en el contexto nacional, promoviendo una respuesta más eficiente y oportuna frente a emergencias en salud pública (17), (18).Adicionalmente, en el desarrollo del proyecto se buscó dar respuesta a las siguientes subpreguntas:

- ¿Cuáles fueron (o están siendo) las técnicas más idóneas para preparar, limpiar y organizar los datos abiertos de COVID19 de Bogotá, asegurando su calidad y confiabilidad en la construcción de modelos predictivos?
- ¿Cuáles son las variables críticas que podrían impactar significativamente la precisión y el rendimiento de los modelos de predicción, considerando factores demográficos, de infraestructura y ambientales específicos del contexto local?
- ¿Qué técnicas resultan más oportunas para entrenar y estimar los modelos de aprendizaje automático, estadísticos y matemáticos, utilizando los instrumentos y algoritmos más adecuados al problema y a la disponibilidad de datos en Bogotá?
- ¿Qué criterios de evaluación estandarizados se han definido o se deberían aplicar para comparar la eficacia de cada modelo, considerando métricas como precisión, especificidad, sensibilidad, análisis costo-beneficio y tiempo de respuesta?

1. OBJETIVOS DEL PROYECTO

1.1. OBJETIVO GENERAL

Desarrollar y evaluar modelos predictivos para brotes de enfermedades infecciosas en salud pública, mediante el uso de datos abiertos de COVID19, comparando enfoques de modelación matemática, estadística y aprendizaje automático, con el fin de identificar el modelo más eficaz para anticipar brotes y apoyar intervenciones tempranas en contextos urbanos como el de Bogotá.

1.2. OBJETIVOS ESPECÍFICOS

- **Preparar, limpiar y organizar datos abiertos de COVID19 disponibles para la ciudad de Bogotá** para asegurar su calidad y confiabilidad en la construcción de modelos predictivos de brotes de salud pública.
- **Identificar las variables más relevantes** que inciden en la aparición y propagación de brotes en el contexto local de Bogotá, considerando aspectos demográficos, de infraestructura en salud y factores ambientales.
- **Entrenar, estimar y ajustar modelos predictivos** basados en enfoques automáticos, estadísticos y matemáticos utilizando las técnicas y algoritmos más adecuados al problema y a los datos y el comportamiento de la serie temporal
- **Comparar el desempeño de los modelos desarrollados** mediante métricas estandarizadas de precisión, sensibilidad, especificidad, tiempo de respuesta y análisis costo-beneficio, con el fin de determinar qué enfoque es más eficaz para intervenciones tempranas en salud pública en Bogotá.

2. MARCO TEÓRICO Y ANTECEDENTES

2.1. MARCO TEÓRICO

La enfermedad por coronavirus 2019 (COVID19) es una patología infecciosa causada por el virus SARS-CoV-2, identificado por primera vez a finales de 2019. Desde su aparición, esta enfermedad se propagó rápidamente a nivel mundial, dando lugar a una pandemia que representó uno de los mayores retos sanitarios del siglo XXI. Su alta transmisibilidad, combinada con la ausencia inicial de tratamientos específicos y vacunas, generó un incremento abrupto en la demanda de servicios de salud, poniendo en evidencia debilidades estructurales en los sistemas sanitarios a nivel global (3), (4).

Desde el punto de vista clínico, el COVID19 presenta un espectro amplio de manifestaciones, que van desde infecciones asintomáticas hasta cuadros graves de insuficiencia respiratoria aguda. Los síntomas más comunes incluyen fiebre, tos seca, fatiga, dificultad respiratoria y pérdida del gusto y del olfato, aunque en casos severos pueden presentarse complicaciones sistémicas como neumonía, trombosis, falla multiorgánica y, en algunos casos, la muerte (3). Esta variabilidad clínica dificultó la detección temprana de casos e incrementó la complejidad en la gestión epidemiológica de la enfermedad.

A nivel mundial, el tratamiento del COVID19 principalmente estuvo orientado al manejo sintomático y al soporte clínico, especialmente en pacientes hospitalizados. Durante las primeras fases de la pandemia, la ausencia de terapias antivirales específicas llevó a la implementación de protocolos basados en oxigenoterapia, ventilación mecánica y el uso controlado de medicamentos antiinflamatorios y anticoagulantes en casos graves (4), (8). Con el avance de la investigación científica, se desarrollaron tratamientos más específicos y se estandarizaron guías clínicas, lo que permitió reducir la letalidad en etapas posteriores de la pandemia.

Uno de los hitos más relevantes en la lucha contra el COVID19 fue el desarrollo y la implementación masiva de vacunas. En un tiempo sin precedentes, diferentes plataformas tecnológicas como vacunas de ARNm, vectores virales y virus inactivados demostraron ser efectivas para reducir la gravedad de la enfermedad, las hospitalizaciones y la mortalidad (20). No obstante, la efectividad de las campañas de vacunación dependió en gran medida de factores logísticos, sociales y de aceptación por parte de la población, lo que generó diferencias significativas en los resultados entre países y regiones.

En el contexto latinoamericano, y particularmente en Colombia, el COVID19 tuvo un impacto considerable tanto en términos sanitarios como sociales. El país enfrentó múltiples olas de contagio, caracterizadas por picos abruptos en el número de casos y fallecimientos, lo que tensionó la capacidad hospitalaria y evidenció desigualdades en el acceso a los servicios de salud. Estas dinámicas resaltaron la importancia de contar con herramientas de predicción que permitieran anticipar escenarios críticos y apoyar la toma de decisiones en salud pública (21), (12).

Dentro del territorio nacional, la ciudad de Bogotá se consolidó como uno de los principales epicentros de la pandemia. Su alta densidad poblacional, la intensa movilidad urbana y la concentración de actividades económicas favorecieron la rápida propagación del virus. Además, Bogotá concentró una proporción significativa de los casos confirmados y de las hospitalizaciones, convirtiéndose en un

escenario clave para el análisis de la dinámica temporal del COVID19 y la evaluación de estrategias de control y mitigación (14), (22).

En este contexto urbano, la gestión del COVID19 implicó desafíos adicionales, como la asignación eficiente de camas de unidades de cuidados intensivos, la implementación de medidas de restricción de movilidad y el seguimiento continuo de indicadores epidemiológicos. Estas condiciones hacen de Bogotá un caso de estudio relevante para el desarrollo y comparación de modelos predictivos, ya que permite evaluar su capacidad para capturar patrones complejos de transmisión en entornos altamente dinámicos y heterogéneos.

Desde una perspectiva de ciencia de datos, el análisis del COVID19 en Bogotá ofrece una oportunidad para integrar información epidemiológica, demográfica y temporal en modelos matemáticos, estadísticos y de aprendizaje automático. La comparación de estos enfoques permite no solo mejorar la comprensión de la evolución de la enfermedad, sino también identificar herramientas con mayor potencial para apoyar la planificación sanitaria y la respuesta ante futuros brotes. De este modo, el estudio del COVID19 se convierte en un eje central para fortalecer la vigilancia epidemiológica y avanzar hacia una salud pública más predictiva y resiliente (9), (17), (16).

En respuesta a la complejidad inherente a la dinámica de propagación de las enfermedades infecciosas, y particularmente del COVID19, la modelación epidemiológica se ha consolidado como una herramienta fundamental para el análisis, la comprensión y la predicción de brotes. La diversidad de factores que intervienen en estos procesos biológicos, sociales, demográficos y temporales ha impulsado el desarrollo de distintos enfoques de modelación, entre los que se destacan los modelos matemáticos, los modelos estadísticos y las técnicas de aprendizaje automático. Cada uno de estos enfoques aborda el fenómeno desde perspectivas complementarias: los modelos de machine learning explotan la capacidad computacional para identificar relaciones complejas y no lineales; los modelos estadísticos se enfocan en el análisis de patrones y tendencias históricas de los datos y los modelos matemáticos permiten representar la dinámica de transmisión mediante ecuaciones que describen el comportamiento del sistema. La comparación de estos enfoques resulta clave para evaluar su desempeño, interpretabilidad y aplicabilidad en contextos reales de salud pública, como el de la ciudad de Bogotá, donde la anticipación de brotes constituye un elemento estratégico para la toma de decisiones sanitarias.

2.1.1. Modelos Matemáticos

Los modelos matemáticos desempeñan un papel fundamental en la salud pública, ya que permiten comprender, anticipar y controlar la propagación de enfermedades infecciosas. Su importancia radica en que transforman datos clínicos y epidemiológicos en herramientas de análisis que apoyan la toma de decisiones estratégicas y operativas. Los modelos matemáticos utilizados en la predicción de brotes en salud pública incluyen sistemas dinámicos, ecuaciones diferenciales y métodos de simulación (7). Estos modelos permiten predecir el comportamiento futuro de una enfermedad, estimando variables clave como el número de casos esperados, picos de infección y duración del brote, esto es vital para preparar los sistemas de salud, gestionar recursos y aplicar medidas de control a tiempo (vacunación, cuarentenas, campañas de prevención). Los modelos matemáticos son fundamentales para comprender

y predecir los mecanismos de propagación de una epidemia porque ayudan a pronosticar brotes importantes, a detectar patrones y monitorear características que pueden sugerir medidas adecuadas para controlar la propagación de enfermedades (4).

El modelo SEIR

Los modelos epidemiológicos matemáticos son modelos compartimentales deterministas, se basan en dividir a la población en grupos o compartimentos según el estado de infección y modelar los flujos entre ellos usando ecuaciones, es decir, son modelos matemáticos que representan la evolución de una enfermedad en una población, categorizando a los individuos en estados mutuamente excluyentes, esto es:

- S: Susceptibles (pueden contraer la enfermedad)
- E: Expuestos (infectados, pero aún no transmiten la infección)
- I: Infectados (pueden transmitir la infección)
- R: Recuperados (adquieren inmunidad temporal o permanente)

Dentro de los modelos clásicos está el modelo SIR (Susceptible – Infectado – Recuperado) ideal para enfermedades que otorgan inmunidad permanente una vez superadas, el modelo SIS (Susceptible – Infectado – Susceptible) especial para enfermedades que no generan inmunidad después de la recuperación, el Modelo SIRS (Susceptible – Infectado – Recuperado – Susceptible) ideal para enfermedades con inmunidad temporal y el modelo SEIR (Susceptible – Expuesto – Infectado – Recuperado) es ideal para enfermedades con periodo de incubación o latencia. Para el presente trabajo se ajusta un modelo SEIR basado en el estado epidemiológico de los individuos pues mejora la precisión en enfermedades donde las personas no son infecciosas inmediatamente después del contagio. Los procesos de transmisión de agentes infecciosos en poblaciones hospederas representan uno de los principales objetivos de estudio de los modelos epidemiológicos compartimentales, los cuales se basan en sistemas de ecuaciones diferenciales para el movimiento de la población a través de estados discretos (14).

El modelo SEIR es una herramienta fundamental en epidemiología matemática y salud pública pues permite simular la dinámica de transmisión de enfermedades infecciosas con período de incubación, como la influenza, el ébola o el COVID19. El modelo ayuda a anticipar cómo evolucionará una epidemia en el tiempo, estimando cuándo alcanzará su punto máximo, cuántas personas serán infectadas, y cuánto durará. Esta predicción es esencial para planificar respuestas sanitarias. los modelos de tipo SEIR ayudan a prever la evolución de una enfermedad, obtener estimaciones de sus características, como las tasas de mortalidad y hospitalización, y conocer el impacto de las intervenciones para predecir la evolución de COVID19. (15).

En este modelo, el flujo de un grupo a otro es de la siguiente manera:

$$S \rightarrow E \longrightarrow I \rightarrow R \qquad (1)$$

La población total $N\,(t)$ es la suma de todos los grupos mutuamente excluyentes, esto es:

$$N = S(t) + E(t) + I(t) + R(t) \quad (2)$$

El sistema de ecuaciones diferenciales ordinarias (EDO) del modelo SEIR es el siguiente:

$$\begin{aligned} \frac{dS}{dt} &= \mu(N - S) - \beta\frac{(SI)}{N} - vS \\ \frac{dE}{dt} &= \beta\frac{(SI)}{N} - (\mu + \sigma)E \\ \frac{dI}{dt} &= \sigma E - (\gamma + \mu)I \\ \frac{dR}{dt} &= \gamma I - \mu R + vS \end{aligned} \quad (3)$$

Aquí los parámetros están definidos de la siguiente manera:

- $\beta = tasa\ de\ transmisión$
- $\sigma = tasa\ expuesta - infectada$
- $\gamma = tasa\ de\ recuperación$
- $\mu = tasa\ demográfica$
- $v = tasa\ de\ vacunación$

El sistema está compuesto por cuatro ecuaciones diferenciales; cada una de ellas describe el comportamiento de los individuos susceptibles, expuestos, infectados y recuperados. La ecuación de susceptibles se divide en tres términos, el primer término $\mu(N - S)$ refleja que los susceptibles aumentan por nacimiento y disminuyen por muertes naturales, el segundo término $\beta\frac{(SI)}{N}$ representa los susceptibles que se contagian y el tercer término vS son los susceptibles vacunados por unidad de tiempo y adquieren inmunidad. La ecuación de expuestos describe la dinámica de personas infectadas, pero no contagiosas, el término $\beta\frac{(SI)}{N}$ son exactamente los que salen de S por infección y el término $(\mu + \sigma)E$ son los expuestos que mueren por causas no relacionadas y los expuestos que pasan a infectados. En la ecuación de infectados describe la dinámica de los que sí transmiten la enfermedad, el término σE son los expuestos que completan incubación y el término $(\gamma + \mu)I$ son los infectados que se recuperan y los infectados que mueren por causas demográficas. La última ecuación pertenece a los recuperados que describe la población protegida contra la infección, el término γI son los infectados que superan la enfermedad, el término μR son los recuperados que mueren naturalmente y el término vS son los susceptibles vacunados que entran a R. En resumen S cambian por nacimientos, infecciones y vacunación, E aumentan por contagio, disminuyen por incubación y muertes, I aumentan por incubación, disminuyen por recuperación y muertes, y R aumentan por recuperación y vacunación, disminuyen por muertes.

En este modelo se relaciona el número básico reproductivo que se define de la siguiente manera:

$$R_0 = \frac{\sigma\beta}{(\sigma + \mu)(\gamma + \mu)} \quad (4)$$

Esta ecuación representa el número promedio de infecciones secundarias que produce un individuo infectado durante todo su periodo infeccioso, en una población completamente susceptible. Esta ecuación se puede interpretar como:

$$R_0 = \frac{\beta}{\gamma + \mu} \quad x \quad \frac{\sigma}{\sigma + \mu} \tag{5}$$

Donde R_0 es la multiplicación de los contagios durante la fase infecciosa por la probabilidad de llegar a infectar. Si el valor de $R_0 > 1$ la epidemia crece, si $R_0 = 1$ La epidemia se mantiene; en caso contrario, la epidemia desaparece.

2.1.2. Modelos Estadísticos

Los modelos estadísticos en salud pública permiten analizar la distribución y los determinantes de las enfermedades en las poblaciones (15). A través de técnicas estadísticas clásicas, se pueden generar predicciones basadas en observaciones pasadas y suposiciones sobre el comportamiento futuro de las variables de interés.

En el ámbito de la salud pública, entre los modelos estadísticos más utilizados se encuentra la regresión, que es una técnica estadística utilizada para predecir una variable dependiente a partir de una o más variables independientes. La regresión logística y la regresión lineal son ampliamente utilizadas para predecir el comportamiento de la propagación de enfermedades.

Los modelos de series temporales permiten identificar patrones de tendencia, estacionalidad y dependencia temporal, facilitando la anticipación de escenarios críticos y apoyando la toma de decisiones. Entre los modelos estadísticos más utilizados se encuentran los modelos de la familia ARIMA, ampliamente aplicados en epidemiología, economía y ciencias sociales, son útiles para prever el comportamiento futuro de los brotes de enfermedades, los SARIMA (*Seasonal ARIMA*) extiende ARIMA para capturar estacionalidad explícita en la serie temporal SARIMA y SARIMAX (*Seasonal ARIMA with eXogenous variables*) incorpora variables exógenas al marco SARIMA.

Modelo ARIMA

Los modelos ARIMA (*AutoRegressive Integrated Moving Average*) es un tipo de modelo estadístico paramétrico diseñado para describir y predecir series temporales no estacionarias, mediante la combinación de tres componentes: Autorregresivo (AR), Integrado (I) y Media móvil (MA). Su formulación general es:

$$ARIMA(p,d,q): \left(1 - \sum_{i=1}^{p} \phi_i L_i\right)(1-L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\varepsilon_t \tag{6}$$

- $p = Términos\ autorregresivos.$
- $d = Diferenciaciones\ para\ estacionariedad.$

- $q = Términos\ de\ media\ móvil.$

Si los datos muestran estacionalidad (ej. picos anuales de dengue), se usan modelos de tipo ARIMA:

$$ARIMA(p,d,q)(P,D,Q)s \quad (7)$$

En esto, un primer estudio indicó que, en la ciudad de Cali en el año 2021, fue posible aplicar el modelo ARIMA$(1,1,1)$ $(1,1,1)$ y se pudo lograr la predicción del dengue, logrando un RMSE (*Root mean square error*), utilizado como métrica para medir la diferencia promedio entre los valores predichos del modelo y los observados, el cual fue de 18,5 frente a modelos lineales simples con un RMSE de 25,3 (error cuadrático medio).

Ecuación general del modelo ARIMA

Después de aplicar d diferenciaciones a la serie original y_t , el modelo se expresa como:

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t \quad (8)$$

donde:

- B el operador rezago $\left(B_{yt} = y_{t-1}\right)$
- $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ \cdots \ - \phi_p B^p$
- $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ \cdots \ + \theta_q B^q$
- ε_t Es un ruido blanco con media cero y varianza constante, es decir, que el modelo ha capturado toda la información predecible de los datos, y lo que queda (residuos) es aleatorio, sin patrones ni correlaciones.

ARIMA trabaja bajo los siguientes supuestos, asumiendo que:

- La serie puede hacerse estacionaria mediante diferenciación.
- Los residuos del modelo son ruido blanco.
- La relación entre valores pasados es lineal.
- No existen cambios estructurales abruptos en el período analizado.

Modelo SARIMA

El modelo SARIMA (*Seasonal ARIMA*) extiende ARIMA para capturar estacionalidad explícita en la serie temporal. Es especialmente útil cuando el comportamiento de la serie se repite de manera periódica, semanal o anual, por ejemplo.

Se representa como:

$$SARIMA(p,d,q)(P,D,Q)s \quad (9)$$

donde:

- $P\ D\ Q$ son las órdenes estacionales.
- s es el período estacional (por ejemplo, s=7 para estacionalidad semanal)

Ecuación general del modelo SARIMA

$$\phi(B)\Phi(B^s)(1-B)^d\ (1-B^s)^D y_t = \theta(B)\Phi(B^s)\varepsilon_t \quad (10)$$

donde:

- $\Phi(B^s)$ y $\Phi(B^s)$ representan los componentes estacionales AR y MA.
- D indica la diferenciación estacional.

SARIMA se ha utilizado para predecir la incidencia de influenza, modelar patrones semanales de reporte de casos y analizar enfermedades con ciclos estacionales, y en este, en especial, porque mejora el ajuste frente a ARIMA en series periódicas.

Modelo SARIMAX

Se expresa como:

$$SARIMAX(p,d,q)(P,D,Q)s + X \quad (11)$$

Ecuación general del modelo SARIMAX

$$y_t = X_t\beta + u_t \quad (12)$$
$$u_t\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D = \theta(B)\Theta(B^s)\varepsilon_t \quad (13)$$

donde:

- X_t es la matriz de variables exógenas.
- β es el vector de coeficientes asociados.
- u_t sigue una estructura SARIMA.

Aunque los modelos ARIMA, SARIMA y SARIMAX han sido ampliamente utilizados, su aplicación a brotes infecciosos presenta desafíos importantes:

- Los datos epidemiológicos suelen no ser estacionarios.
- Existen cambios estructurales abruptos.
- Los conteos de casos no siguen una distribución normal.
- La dinámica depende de múltiples factores externos no observables.

Obtención de hiperparámetros en modelos ARIMA y SARIMA

La selección de los hiperparámetros en los modelos de series de tiempo ARIMA y SARIMA sigue un enfoque metodológico mixto, en el que se pueden combinar el método clásico de *Box–Jenkins* con una búsqueda sistemática tipo *Grid Search*, con el objetivo de garantizar tanto la validez estadística del modelo como su capacidad predictiva.

- **Enfoque Box–Jenkins**

El método de **Box–Jenkins** constituye una formación iterativa de identificación, estimación y validación de modelos ARIMA – SARIMA.

El proceso iterativo de identificación se centra en los siguientes pasos:

1. **d (diferenciación)**: Determina el orden de diferenciación necesario para hacer la serie estacionaria, entonces Si el p-value resulta ser superior al 0.05, se aplica diferenciación a la serie o hasta obtener una prueba p-value inferior al mismo.

2. **p (AR)**: Usa el ACF (*Autocorrelation Function*) para identificar el orden del AR. Si ACF decae lentamente, entonces necesita diferenciación, por el contrario, si ACF tiene cortes bruscos después del lag k, entonces MA(q) con q=k.

3. **q (MA)**: Usar el PACF (*Partial Autocorrelation Function*) para identificar el orden MA. Si el PACF decae lentamente, entonces necesita diferenciación. Si el PACF tiene cortes bruscos después del lag k, entonces, AR(p) con p=k.

El proceso de estimación de los parámetros del modelo se realiza bajo el enfoque de máxima verosimilitud y, para la validación de los residuos, estos deben ser de ruido blanco a través de la prueba de Ljung-Box para autocorrelación en residuos, la cual, si un *p-value* resulta ser superior al 0.05, entonces indica residuos de ruido blanco.

- **Enfoque Grid Search**

Busca automáticamente la mejor combinación de parámetros minimizando criterios como AIC o BIC. Una vez definidos los rangos plausibles mediante Box–Jenkins, se implementa una búsqueda exhaustiva tipo *Grid Search* sobre las combinaciones posibles de parámetros.

Para cada combinación de hiperparámetros se estima el modelo y se evalúa utilizando:

AIC (Akaike Information Criterion)
BIC (Bayesian Information Criterion)

El modelo óptimo seleccionado es aquel que minimiza el AIC (y/o BIC), presenta convergencia numérica estable, cumple con los supuestos del modelo y el enfoque garantiza un equilibrio entre calidad del ajuste y parsimonia del modelo.

En síntesis, los modelos ARIMA, SARIMA y SARIMAX constituyen herramientas estadísticas valiosas para el análisis de series temporales; sin embargo, su efectividad depende del cumplimiento de sus supuestos y del contexto de aplicación. En el ámbito de la salud pública, su uso debe evaluarse cuidadosamente, priorizando no solo la precisión estadística, sino también la estabilidad, interpretabilidad y viabilidad operativa de los modelos seleccionados.

2.1.3. Modelos de Aprendizaje Automático (Machine Learning)

El aprendizaje automático (ML) es una subdisciplina de la inteligencia artificial que se enfoca en el desarrollo de algoritmos capaces de aprender y hacer predicciones o decisiones basadas en datos (23). Los modelos de aprendizaje automático son herramientas poderosas para la predicción de brotes en salud pública debido a su capacidad para identificar patrones complejos en datos masivos y no estructurados, como registros de salud, datos meteorológicos, información sobre movilidad de la población, etc.

Los modelos de aprendizaje automático se utilizan en la predicción de brotes de enfermedades, por ejemplo, utilizando datos históricos de casos de enfermedades, variables meteorológicas y de movilidad para anticipar el inicio y la propagación de brotes como la gripe, el dengue, el COVID19, entre otros.

La predicción de brotes en salud pública es un desafío multidisciplinario que requiere el uso de diversas herramientas analíticas y de modelación. Los modelos de aprendizaje automático, estadísticos y matemáticos tienen sus fortalezas y limitaciones, pero en conjunto pueden ofrecer un enfoque robusto para abordar este desafío. La clave radica en seleccionar y combinar los modelos adecuados según el tipo de datos disponibles, el objetivo específico del análisis y las características del brote en cuestión.

La comparación de estos modelos permite identificar cuál es el más adecuado en función de la precisión, la interpretabilidad y la capacidad de generalizar en escenarios futuros. Mediante la integración de modelos de distintas naturalezas, es posible obtener predicciones más confiables y precisas, lo que, a su vez, contribuye a mejorar la respuesta ante emergencias de salud pública.

XGBoost (EXtreme Gradient Boosting)

XGBoost es un algoritmo basado en *boosting* que ha demostrado un rendimiento sobresaliente en tareas de regresión y clasificación, especialmente cuando se dispone de múltiples variables predictoras. En el contexto de series temporales, *XGBoost* no modela directamente la dependencia temporal, pero al incorporar variables de lag, medias móviles, estacionalidad y otras variables exógenas (edad, ubicación, sexo, etc.), se vuelve altamente efectivo. *XGBoost* optimiza una función objetivo de la forma:

$$L(\phi) = \sum_{i=1}^{n} l\left(y_i, \hat{} y_i^{(t)}\right) + \sum_{k=1}^{t} \Omega\left(f_k\right) \qquad (14)$$

Donde:

- $l(y_i, \hat{} y_i^{(t)})$ Es la función de pérdida (por ejemplo, el error cuadrático medio)

- $f_k \in F$ representa un árbol de regresión
- $\Omega(f_k) = \gamma T + \frac{1}{2}\sum_{j=1}^{T} w_j^2$ es un término de regularización, con T nodos hoja y w_j sus pesos.

Cada nuevo árbol se entrena para minimizar el gradiente de la función de pérdida respecto a la predicción previa.

Entre las principales características de este modelo tenemos: la utilización de *boosting,* donde se ensamblan árboles secuenciales que corrigen errores de los anteriores; se permite la inclusión de múltiples variables exógenas y temporales, así como la ingeniería de características*;* además, no requiere que la serie sea estacionaria.

Redes Neuronales Recurrentes (LSTM / GRU)

Las redes neuronales recurrentes (RNN) están diseñadas para procesar secuencias de datos, lo cual las hace especialmente aptas para modelar series temporales. A diferencia de modelos basados en árboles, las RNN modelan explícitamente la dependencia temporal, conservando información pasada mediante estados internos. Las variantes LSTM (*Long Short-Term Memory*) y GRU (*Gated Recurrent Unit*) resuelven el problema del desvanecimiento del gradiente, permitiendo capturar dependencias a largo plazo. Una RNN tradicional actualiza su estado oculto con:

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1} + \ b_h) \tag{15}$$

Las versiones LSTM y GRU incorporan puertas para controlar el flujo de información.

LSTM:

$$\begin{gathered} f_t = \sigma\big(W_f \,.\, [h_{t-1}, x_t] + b_f\big) \quad (puerta\ de\ olvido) \\ i_t = \sigma(W_i \,.\, [h_{t-1}\,, x_t] + b_i) \quad (puerta\ de\ entrada) \\ o_t = \sigma(W_o \,.\, [h_{t-1}, x_t] + b_o) \quad (puerta\ de\ salida) \\ c_t = \tanh(W_c \,.\, [h_{t-1}\,, x_t] + b_c) \quad (estado\ candidato) \\ c_t = f_t \odot c_{t-1} + i_t \odot\tilde{} c_t \\ h_t = o_t \odot \tanh\,(c_t) \end{gathered} \tag{16}$$

GRU (simplificado):

$$\begin{gathered} z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (puerta\ de\ actualización) \\ r_t = \sigma(W_r x_r + U_r h_{t-1}) \quad (puerta\ de\ reinicio) \\ h_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) \\ h_t = (1 - z_t) \odot h_{t-1} + z_t \odot\tilde{} h_t \end{gathered} \tag{17}$$

Estas redes han demostrado ventajas en el modelamiento de series temporales, ya que modelan secuencias temporalmente correlacionadas, capturan relaciones no lineales y dependencias de largo plazo. Además, soportan múltiples entradas, como variables exógenas multivariadas. Aunque su modelado es computacionalmente costoso y generalmente requiere normalización de datos y mayor

ajuste de hiperparámetros.

2.2. ANTECEDENTES

A continuación, se mencionarán algunos proyectos que trabajaron con modelos de predicción de COVID19 y otros brotes de enfermedades en salud pública. Para esto se realizó una búsqueda en *Google Scholar* con las palabras clave de “modelos estadísticos”, “modelos matemáticos”, “Machine learning”, "predicción" y "COVID19", con un límite de publicación no mayor a 5 años, es decir, entre los años 2019–2024.

En el año 2024 se realizó un proyecto llamado “*Machine learning in predicting severe acute respiratory infection outbreaks*” (16) donde se determinan que en Brasil anualmente se producen brotes de síndrome respiratorio agudo severo (SARS), con picos estacionales que varían según las regiones geográficas. Este estudio tuvo como objetivo predecir los brotes de IRAG (Infección Respiratoria Aguda Grave) basándose en modelos generados con aprendizaje automático utilizando datos de notificación de hospitalizaciones por IRAG. En este estudio, se utilizaron datos de la notificación de casos de hospitalización por IRAG en Brasil de 2013 a 2020, excluyendo los casos de IRAG causados por COVID19. Observaron una correlación entre la ocurrencia de IRAG y los períodos lluviosos y fríos, especialmente en el Sur.

Dado que este estudio trata de series temporales multivariadas con estacionalidad, eligieron la red neuronal recurrente (RNN) LSTM, para la generación de modelos predictivos se simuló utilizando algoritmos de aprendizaje automático como *Random Forest*, *Naive Bayes, Tree Assemble y el Perceptrón Multicapa (MLP) de Resilient Backpropagation (RPROP)*. Para los modelos predictivos probados, el conjunto de datos se restringió a las siguientes variables: número de notificaciones diarias (utilizado como variable dependiente); fecha de los primeros síntomas; sexo (masculino, femenino); grupo de edad (joven, adulto, adulto mayor); y diagnóstico positivo para IAV o IBV, además, se agregaron al conjunto de datos los promedios diarios de temperaturas mínimas y máximas y la amplitud térmica por región, así como el promedio total de temperaturas por región, los entornos de programación utilizados fueron *pipeline Knime* y R. Dentro de los resultados encontraron que la aplicación del LSTM tuvo una mayor precisión que el uso del método de Promedio Móvil Integrado Autorregresivo Estacional (SARIMA), teniendo en cuenta que no encontraron mejoras significativas en la precisión al utilizar variables de temperatura para generar los modelos predictivos, resaltan que utilizando algoritmos de series temporales, los modelos de predicción generados con redes neuronales mostraron buenos resultados en la predicción de temporadas de brotes de IRAG.

En el año 2022, Garrido & otros (8) realizaron un trabajo llamado “*Modelo matemático optimizado para la predicción y planificación de la asistencia sanitaria por la COVID19”* donde su objetivo principal era desarrollar un modelo matemático diseñado para optimizar las predicciones relacionadas con las necesidades de hospitalización e ingresos en UCI por la COVID19. Para ello utilizaron pacientes de COVID19 hospitalizados, ingresados en UCI, recuperados y fallecidos desde el 15 de marzo hasta el 22 de septiembre del 2020 e implementaron un modelo susceptible, expuesto, infectado y recuperado (SEIR) diseñado específicamente para describir la dinámica de la epidemia a nivel poblacional y a nivel del circuito hospitalario en relación con los pacientes de COVID19. Generaron 3 escenarios: el escenario

inicial representó la evolución predicha, tomando en cuenta las restricciones establecidas; los otros dos escenarios adicionales fueron escogidos entre muchas simulaciones, ya que permiten analizar los efectos de la dilatación temporal de las medidas de contención y el establecimiento de diferentes períodos de restricciones.

En el estudio, la fase de calibración y validación del modelo mostró la validez de las predicciones proporcionadas tras comparar los casos esperados y registrados de hospitalizaciones e ingresos en UCI durante el período. Los autores concluyen que el modelo matemático es capaz de proporcionar predicciones sobre la evolución de la COVID19 con suficiente antelación como para poder conjugar los picos de prevalencia y de necesidades de asistencia hospitalaria y de UCI, con la aparición de ventanas temporales que posibiliten la atención de enfermos no-COVID19. Además, el modelo de previsión epidemiológica que planteamos nos permite evaluar el impacto de las distintas estrategias de restricción poblacional frente a la COVID19, considerando su duración, intensidad y el contexto basal de incidencia y prevalencia, así como prever el nivel de presión asistencial relativo al número de pacientes hospitalizados e ingresados en UCI.

En el año 2021, Oropesa Fernández & otros (24) realizaron un estudio llamado "Modelo estadístico para estimar el impacto histórico de la influenza sobre la mortalidad en Cuba" donde el objetivo general fue de estimar el impacto histórico de la influenza tipo A y B y los subtipos A(H3N2) y A(H1N1) sobre la mortalidad mediante el ajuste de un modelo de regresión a las condiciones estacionales específicas de Cuba. Acá subrayan que las estadísticas confirman cada año entre 3 y 5 millones de enfermos y alrededor de 250 mil a 500 mil fallecidos por esta causa, con un mayor porcentaje entre los individuos de mayor riesgo. En este estudio se ejecutó un estudio longitudinal y retrospectivo. En un primer paso se ajustaron dos modelos de Poisson con la mortalidad por influenza y neumonía total y las personas mayores o iguales a los 65 años como variables respuesta en los cinco meses de mayor positividad en influenza, desde la temporada 1987-1988 hasta la 2004-2005, y los positivos en tipo A y en tipo B como explicativas. En otro par de modelos se estimó el impacto del A(H3N2) y del A(H1N1), considerando como respuesta a los fallecidos atribuidos previamente al tipo A.

En los resultados encontraron que se atribuyeron a la influenza 7803 fallecidos entre todas las edades y 6152 entre las personas ≥ 65 años, con un 56,3 % asociados al A(H3N2), el 17,6 % al A(H1N1) y el 26,1 % al tipo B. Además, afirman que las tasas de mortalidad influenza y neumonía total atribuidas a la influenza en este trabajo (4,0 por 100 000 habitantes) y en las personas de 65 años o más (33,6 por 100 000 habitantes), son coherentes con las obtenidas en Singapur (2,9 y 46,9) y con las de Hong Kong (4,1 y 39,3) calculadas mediante una regresión de Poisson. Resaltan que el valor del modelo ajustado se evidencia en la contraposición de las condiciones ecológicas de Cuba frente a la bien definida estacionalidad de la influenza y su correspondencia con la mortalidad, propia de los países templados, donde, aún con esta condición, se considera difícil cuantificar sus magnitudes relacionales. Fernández & otros señalan que con esto se demuestra la posibilidad de ajustar estos modelos de regresión a otros virus respiratorios y a la actual pandemia por la COVID19, en las condiciones estacionales de Cuba.

Badal (20) en el año 2021 realiza un estudio llamado "*Modelo de predicción de riesgo hospitalario por COVID19 y su aplicación en la evaluación de estrategias de vacunación*" cuyo objetivo fue desarrollar un modelo matemático que pronostique la probabilidad de manifestar un cuadro severo de SARS-CoV-2 a

partir de datos médicos desagregados que es posible conseguir antes de que los pacientes se infecten, con el propósito de evaluar estrategias para la organización del proceso de vacunación en Chile. Para ello desarrollaron un modelo de predicción de riesgo por COVID19 en base a datos demográficos y condiciones preexistentes de los infectados con COVID19 en México. Se entrenaron cuatro algoritmos de clasificación: Random Forest, *XGBoost*, Regresión por Mínimos Cuadrados Parciales (PLS-DA) y Regresión Logística, con tres estructuras de modelos —una incorporó técnicas de selección de variables—, y se compararon a partir del valor de sus métricas de desempeño. El modelo de *XGBoost* con selección de variables alcanzó la mayor capacidad discriminativa (AUC = 0,82, sensibilidad ≈ 50%) que es comparable con investigaciones anteriores, por lo que, fue seleccionado para ser aplicado en la evaluación de estrategias de vacunación. Los resultados de este estudio indican que es factible obtener un rendimiento comparable a otros estudios predictivos de riesgo por COVID19, utilizando únicamente información previa a la infección, como lo son las comorbilidades y la demografía. Asimismo, el estudio demuestra que algunas comorbilidades, la edad y el sexo de los pacientes son variables relevantes para predecir un cuadro severo. Los resultados del análisis exploratorio y la evaluación de la importancia de las variables al predecir coinciden con los hallazgos de investigaciones médicas previas que han reportado que la edad y las enfermedades subyacentes (tales como hipertensión, diabetes, y enfermedad cardiovascular) pueden ser factores de riesgo para los pacientes de COVID19. En el trabajo se resalta que el modelo permite clasificar a los pacientes según una escala de riesgo, la cual se asocia a sus características intrínsecas y anteriores a la infección. De esta manera, provee información provechosa para identificar a los grupos de mayor riesgo al virus, es decir, que son más propensos a colapsar el sistema de salud.

Prades & Marín en el año 2020 realizan un estudio llamado "*Modelos estadísticos para las predicciones de la COVID19 en Cuba*" (12) cuyo objetivo fue realizar un análisis de modelación estadística combinando 6 modelos de pronóstico para predecir la aparición de casos positivos diarios, activos y fallecidos por COVID19 en Cuba ya que la pandemia ha afectado a comunidades y economías sin precedentes en todo el mundo. En este estudio utilizaron los datos reportados diariamente del 11 de marzo al 25 de mayo publicados en el sitio web Cuba Debate. Los datos utilizados fueron procesados y analizados usando Microsoft Office Excel 2016 y el software estadístico STATGRAPHICS. Centurion. XV versión 15.2.14 con un ajuste de raíz cuadrada a los modelos: A: Promedio móvil simple de 2 términos, B: Suavización exponencial simple con alfa = 0.4888, C: Suavización exponencial de Brown con alfa = 0.235, D: ARIMA (2,0,0) con constante. El modelo E ARIMA (2,0,1) x (1,0,0)14 con constante se ajustó por el método de *Box Cox* con una estacionalidad de 14 y el F: ARIMA (1,1,6) sin ajuste. A estos modelos propuestos se les calculó el desempeño mediante los estadísticos: MAE, RMSE, MAPE y MASE, así como el análisis de residuales.

Dentro de los resultados, encontraron que los modelos A y B dan una tendencia constante de 8 y 9 casos positivos, respectivamente, para el día 22 de julio. El modelo C indica una ligera disminución de los casos con 4 ese mismo día y el modelo D una tendencia al aumento con 19 casos. El modelo E refleja un mínimo de 126 casos el día 3 de junio y luego un aumento de los casos hasta alcanzar el 22 de julio un valor de 374 casos activos hospitalizados. En el modelo F se apreció una tendencia a mantenerse constante el número de fallecidos por encima de 80 casos en la primera quincena de julio. Los autores concluyeron que los datos obtenidos de los modelos predictivos fueron útiles porque proporcionan un pronóstico para la epidemia de COVID19, lo que representa una herramienta válida y objetiva para monitorear el control de infecciones. Además, afirma que todas las instituciones involucradas en la salud pública y el control

de infecciones pueden beneficiarse de estos datos porque al usar estos modelos, pueden construir diariamente un pronóstico confiable para la epidemia de COVID19.

Finalmente, Bezerra & otros (22) en el año 2020 realizaron un estudio llamado *"Estimación y predicción de casos de COVID19 en metrópolis brasileñas"* cuyo objetivo fue estimar la tasa de transmisión, el pico epidemiológico y el número de muertes por el nuevo coronavirus mediante un modelo matemático y epidemiológico de casos susceptibles, infectados y recuperados (SIR) a las nueve capitales brasileñas con mayor número de casos de infección. Se analizaron 2.829 casos confirmados de COVID19 en nueve capitales brasileñas para la solución de las ecuaciones diferenciales para cada uno de los nueve escenarios. Para probar el ajuste del modelo, se utilizó el logaritmo natural del número de casos observado y previsto. Dentro de los resultados, encontraron que las nueve metrópolis estudiadas mostraron una curva ascendente de casos confirmados de COVID19, los datos de predicción apuntan al pico de la infección entre finales de abril y principios de mayo. Fortaleza y Manaus tuvieron las tasas de transmisión más altas, Río de Janeiro puede tener el mayor número de personas infectadas y Florianópolis el menor. Concluyeron que las estimaciones de la tasa de transmisión, el pico epidemiológico y el número de muertes por coronavirus en metrópolis brasileñas presentaron cifras expresivas e importantes que el Ministerio de Salud de Brasil necesita considerar. Los resultados confirman la rápida propagación del virus y su alta mortalidad en el país.

Los proyectos mencionados anteriormente dan prueba de que los modelos matemáticos, estadísticos y de machine learning arrojan muy buenos resultados en cuanto a la predicción de brotes de enfermedades infecciosas como el COVID19. Los modelos matemáticos como los SIR (Susceptibles, Infectados, Recuperados) ayudan a prever la dinámica de transmisión en diferentes escenarios, proporcionando a las autoridades información valiosa para tomar decisiones preventivas. Los modelos estadísticos son capaces de analizar grandes volúmenes de datos epidemiológicos, demográficos y ambientales, siendo claves para la identificación de correlaciones y factores de riesgo asociados a los brotes, como estacionalidad, densidad poblacional o acceso a servicios de salud y el machine learning por medio de redes neuronales permite construir modelos más complejos y específicos que pueden adaptarse a características locales, como las tasas de vacunación o los flujos migratorios. Además, su capacidad de procesamiento masivo lo hace ideal para manejar datos en tiempo real y de múltiples fuentes. Los modelos matemáticos, estadísticos y de machine learning no solo mejoran la capacidad de predicción, sino que también permiten una respuesta más ágil y eficiente frente a brotes en salud pública.

3. PREPARACIÓN, LIMPIEZA Y ORGANIZACIÓN DE DATOS ABIERTOS DE COVID19

3.1. OBTENCIÓN DEL CONJUNTO DE DATOS

El conjunto de datos utilizado en este estudio corresponde a información oficial publicada por el Instituto Nacional de Salud de la República de Colombia (INS), el cual fue descargado del portal de datos abiertos del Gobierno Nacional (datos.gov.co). En particular, se empleó la base de datos disponible en dicho repositorio, que constituye la fuente más veraz, completa y actualizada sobre los casos de COVID19 reportados a nivel nacional. A partir de esta información se extrajo un subconjunto correspondiente a los casos registrados en el Distrito Capital de Bogotá durante el período de la pandemia declarada por el Gobierno Nacional.

La selección de Bogotá, D. C., como unidad de análisis responde a su relevancia a nivel nacional, así como al elevado número de casos reportados y a la complejidad demográfica, social y territorial que caracteriza a la ciudad. Estas condiciones permiten que los resultados obtenidos presenten un mayor grado de generalización, al incorporar un nivel significativo de complejidad tanto en los modelos analíticos empleados como en el fenómeno epidemiológico objeto de estudio.

Adicionalmente, dado que la información fue recolectada, analizada y depurada por una entidad de carácter técnico y especializado, el conjunto de datos no presenta problemas significativos asociados a valores faltantes, registros nulos o duplicados. En consecuencia, las tareas de limpieza, depuración y transformación de los datos no representaron un desafío considerable para el desarrollo del proyecto.

En definitiva, este conjunto de datos permite caracterizar de manera integral el comportamiento de la pandemia de COVID19 en Bogotá, D. C., durante los años 2020, 2021 y 2022. Cabe resaltar que los años 2020 y 2022 no cuentan con información correspondiente a la totalidad de los doce meses, sino únicamente a períodos parciales de estos años. No obstante, el intervalo temporal analizado resulta suficientemente amplio para examinar las distintas fases de la pandemia, así como el comportamiento del fenómeno epidemiológico tanto dentro de cada fase como en los períodos de transición entre ellas, lo que contribuye a una comprensión más profunda de su evolución temporal y de los patrones asociados a su dinámica en el contexto urbano de la ciudad.

3.2. DESCRIPCIÓN GENERAL DEL CONJUNTO DE DATOS

El conjunto de datos analizado, es decir, el subconjunto del conjunto original con los datos únicamente de Bogotá D.C., incluye más de 1,8 millones de registros con información detallada sobre casos individuales confirmados de COVID19. Este incluye las variables de identificación, temporales, geográficas, demográficas, clínicas y epidemiológicas recolectadas por la autoridad sanitaria responsable de la información. Las variables originales correspondientes se registran en la tabla 1.

Tabla 1. Identificación de las variables

Tipo de variable	Nombre original del campo	Tipo de dato	Descripción
Identificación	__id	object	Identificador interno del registro.
	id_de_caso	int64	Identificador único del caso confirmado.
Temporal	fecha_reporte_web	datetime64[ns]	Fecha de publicación del caso en el portal oficial.
	fecha_de_notificacion	object	Fecha de notificación del caso a la autoridad sanitaria.
	fecha_inicio_sintomas	object	Fecha de inicio de los síntomas reportados.
	fecha_diagnostico	object	Fecha de confirmación diagnóstica del caso.
	fecha_recuperado	object	Fecha de recuperación del paciente.
	fecha_muerte	object	Fecha de fallecimiento, en caso de presentarse.
Geográfica	departamento	int64	Código del departamento de notificación.
	departamento_nom	object	Nombre del departamento de notificación.
	ciudad_municipio	int64	Código de la ciudad o municipio de notificación.
	ciudad_municipio_nom	object	Nombre de la ciudad o municipio de notificación.
Demográfica	edad	int64	Edad del paciente.
	unidad_medida	int64	Unidad de medida de la

			edad (años, meses, días).
	sexo	object	Sexo del paciente.
	pertenencia_etnica	float64	Código de pertenencia étnica.
	nombre_grupo	object	Nombre del grupo étnico al que pertenece el
Clínica	estado	object	Estado clínico del paciente (activo, recuperado, fallecido, etc.).
	ubicacion	object	Ubicación del paciente durante el seguimiento (casa, hospital, UCI).
	recuperado	object	Condición de recuperación del paciente.
	tipo_recuperacion	object	Tipo de recuperación (tiempo, PCR, clínica).
Epidemiológica	fuente_tipo_contagio	object	Fuente del contagio (importado, relacionado, en estudio).
	pais_viajo_1_cod	float64	Código del país de procedencia, si aplica.
	pais_viajo_1_nom	object	Nombre del país de procedencia, si aplica.

3.3. DETECCIÓN DE PROBLEMAS Y ACCIONES CORRECTIVAS

3.3.1. Nombres de columnas

En principio, los nombres originales de las columnas no fueron modificados, dado que la mayoría de las variables no serían utilizadas en el análisis de series de tiempo.

3.3.2. Eliminación de columnas

La mayoría de las columnas del conjunto de datos original fueron eliminadas, ya sea por ser redundantes,

innecesarias o simplemente por no resultar relevantes para el análisis específico de series de tiempo que se aborda. En este sentido se procedió de la siguiente forma:

- **Variables de identificación:** Tanto el identificador interno del registro como el del caso confirmado fueron eliminados, ya que no resultan útiles para detectar duplicados ni para el desarrollo de análisis preliminar o de modelado.

- **Variables geográficas:** Fueron eliminadas en su totalidad, dado que perdieron utilidad al filtrar el conjunto de datos original para incluir únicamente los casos de Bogotá, D. C. Al presentar los mismos valores en todo el conjunto de datos, estas columnas no aportan información adicional relevante.

- **Variables epidemiológicas:** Se eliminaron, ya que el estudio no se centra en analizar la procedencia o nacionalidad de los casos, la fuente del contagio ni su incidencia en la dinámica de este.

- **Variables demográficas:** También fueron eliminadas en su totalidad, dado que aportan escasa información para el análisis temporal del contagio y no se consideró necesario ni útil incorporarlas como variables exógenas en el estudio.

- **Variables temporales:** Se eliminaron las variables fecha_reporte_web y fecha_de_notificacion, fecha_muerte, fecha_recuperado y fecha_inicio_sintomas, dado que únicamente proporcionan información de control o seguimiento por parte de la autoridad sanitaria y no resultan relevantes ni útiles para el desarrollo de los modelos.

- **Variables clínicas:** Se eliminaron únicamente las variables ubicación y tipo_recuperación, ya que no se consideran relevantes para los objetivos planteados en el proyecto. No obstante, las demás variables de este grupo se consideraron esenciales para definir el conjunto final destinado al entrenamiento.

3.3.3. Tratamiento de valores nulos

En las variables remanentes, al realizar la búsqueda de valores nulos, se detectaron los siguientes registros faltantes por cada variable (tabla 2).

Tabla 2. Registros faltantes en la Base de datos COVID19

Columna	Valores nulos	Valores únicos
recuperado	8,248	3
fecha_diagnostico	185	1405

.

A estos valores faltantes se les dio el tratamiento conforme a las siguientes reglas:

- **Variable recuperado:** Esta variable es fundamental para el desarrollo del proyecto, dado que, en

conjunto con la variable temporal seleccionada, permitió realizar la agrupación necesaria para definir las series de tiempo objeto de análisis. El número de valores nulos es extremadamente bajo en relación con el tamaño total del conjunto de datos, representando menos del 0,005 % de los registros. En consecuencia, se optó por eliminar estos casos en lugar de aplicar técnicas de imputación.

- **Variable fecha_diagnóstico:** Corresponde a la variable temporal objetivo y, por tanto, constituye la de mayor relevancia dentro del conjunto de datos. Dado que los valores nulos son marginales, estos registros fueron eliminados sin afectar la validez del conjunto de datos para la modelación. Al tratarse de un conjunto de datos a nivel individual, el impacto de esta eliminación sobre los modelos es nulo o imperceptible.

3.3.4. Estandarización de variables categóricas

Para evitar duplicados semánticos y mejorar la limpieza visual, se aplicaron transformaciones a las variables categóricas recuperado y estado:

- Normalización a minúsculas.
- Eliminación de espacios o caracteres en blanco al inicio y al final de las incidencias.

3.3.5. Validación de fechas

Se verificó que las fechas estuvieran dentro de un rango coherente: desde el 6 de marzo de 2020 hasta el 17 de enero de 2024. Se estandarizó el formato para facilitar el análisis temporal.

3.3.6. Duplicados

No se detectaron valores duplicados al ser un conjunto de datos oficial procesado y validado con unicidad de registro.

3.3.7. Agrupación de serie de tiempo

Para la construcción de la serie de tiempo definitiva se procedió de la siguiente manera:

Variable temporal

En primer lugar, se seleccionó como variable temporal a modelar el campo *fecha_diagnóstico*, dado que es el que presenta la información más completa, consistente y confiable dentro del conjunto de datos analizado.

Al respecto, si bien la variable *fecha_inicio_sintomas* podría, en diversos contextos, considerarse un mejor descriptor de la dinámica real del contagio, esta presenta un porcentaje elevado de valores faltantes. En consecuencia, la exclusión de dichos registros del análisis no resulta viable, debido a la pérdida sustancial de información; mientras que la aplicación de técnicas de imputación podría

introducir sesgos significativos que comprometerían la validez de los resultados. Adicionalmente, al tratarse de información reportada directamente por el paciente, esta variable es susceptible a sesgos de recuerdo y a imprecisiones asociadas al olvido o a la percepción subjetiva del individuo al momento de la recolección de los datos.

En contraste, la variable *fecha_diagnóstico* es registrada por profesionales de la salud en un entorno controlado, lo cual contribuye a garantizar, en mayor medida, la calidad, trazabilidad e idoneidad de la información. Por estas razones, se consideró que dicha variable constituye la opción más adecuada para la modelación temporal del fenómeno bajo estudio.

Variable objetivo

La variable a predecir corresponde al número de *infectados* diarios, es una variable creada a partir del recuento de las observaciones o registros diarios En consecuencia, el objetivo será predecir el valor diario de los infectados con COVID19.

Variables endógenas secundarias

Estas variables son de suma importancia, dado que de su suma obtenemos el número de infectados en determinado periodo. Estas son las variables de fallecidos y recuperados. Estas no fueron utilizadas en los modelos como variables predictoras, dada su colinealidad con la variable objetivo.

Variables exógenas

Se incorporaron las siguientes variables exógenas al conjunto de datos ya agrupado:

- **Dosis_vacuna.** Información obtenida del Ministerio de Salud y Protección Social. La información ya se encuentra agrupada diariamente, por lo que se extrae la información y se incorpora al conjunto de datos.

- **Total pasajeros.** Información de tráfico diario de pasajeros en el Aeropuerto Internacional El Dorado. La información en la página web del aeropuerto se encuentra discriminada entre pasajeros en entradas y salidas nacionales e internacionales. Para incorporar la información al conjunto de datos la información se hizo la sumatoria de

- **Temperatura_promedio.** El conjunto de datos de las temperaturas se obtuvo del IDEAM y presenta una granularidad en términos de estación de medida y hora. Este conjunto de datos es extremadamente grande y se agrupó mediante el promedio ponderado de las medidas de la temperatura presentadas en todas las estaciones de la capital en el transcurso del día.

En la tabla 3 se muestra la clasificación de las variables relacionadas y su respectivo rol con la serie.

Tabla 3. Clasificación de variables

Variable	Clasificación principal	Rol en la serie
fecha diagnóstico	Índice temporal	Eje temporal
infectados	Variable endógena / objetivo	Serie principal
fallecidos	Variable endógena secundaria	Resultado
recuperados	Variable endógena secundaria	Estado
dosis_vacuna	Variable exógena	Intervención
total_pasajeros	Variable exógena	Explicativa / movilidad
temperatura_promedio	Variable exógena	Control climática

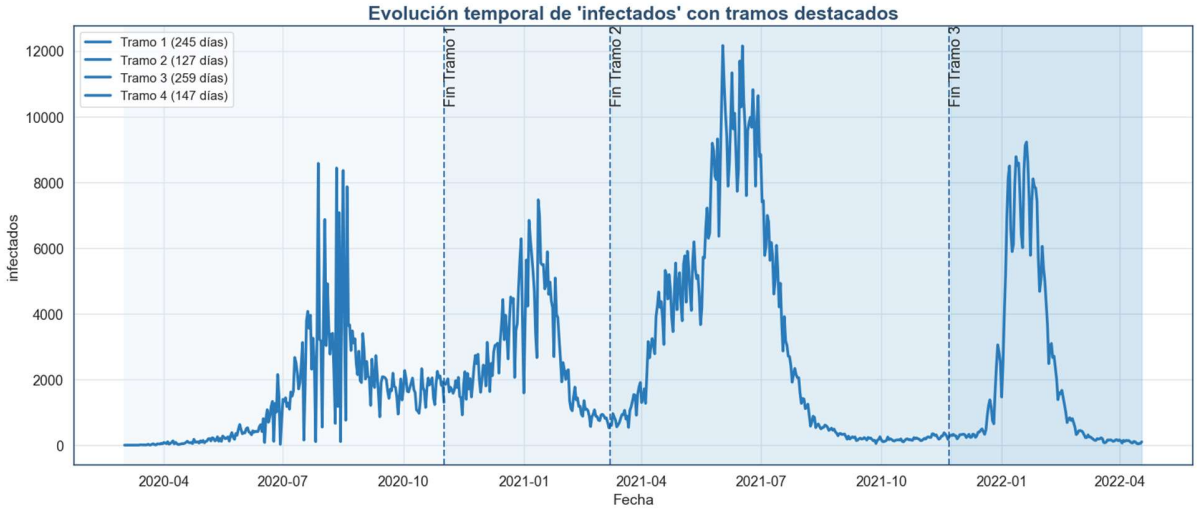
3.3.8. Transformaciones.

En principio, no se efectuó ninguna transformación o escalado sobre las variables, dado que estas fueron definidas en las fases de modelado y entrenamiento, toda vez que cada técnica empleada requiere diferentes tipos de transformación.

3.3.9. Definición de tramos u olas

Con el propósito de representar de manera fiel la evolución epidemiológica del COVID-19 en la ciudad de Bogotá, los modelos fueron definidos y calibrados mediante un fraccionamiento temporal por olas de contagio, claramente identificables en la serie de casos observados (figura 1).

figura 1. Evolución temporal de infectados con tramos destacados



El análisis exploratorio de los datos epidemiológicos de Bogotá evidencia la presencia de picos sucesivos, asociados a cambios en la intensidad de transmisión, la respuesta sanitaria y el comportamiento poblacional, lo que justifica la segmentación de la serie en cuatro olas principales.

Este enfoque reconoce que la dinámica de propagación del SARS-CoV-2 en Bogotá no puede ser descrita

adecuadamente mediante un único conjunto de parámetros constantes a lo largo del tiempo. En concordancia con estudios recientes, la modelación por tramos u olas en la ciudad de Bogotá ha permitido capturar de forma más realista la naturaleza no estacionaria de la epidemia, mejorando la interpretabilidad epidemiológica y el desempeño del modelo frente a enfoques continuos con parámetros invariantes (25). De esta manera, los tramos temporales u olas facilitan la comparación del comportamiento del modelo bajo distintos contextos epidemiológicos propios de la ciudad de Bogotá y garantizan la coherencia metodológica entre los diferentes enfoques de modelación evaluados.

En este mismo sentido, se realizó un filtro teniendo en cuenta la fecha final fijando el periodo de la serie de tiempo completa entre el 02 de marzo de 2020 y el 18 de abril de 2022, correspondiente a 778 observaciones, como se muestra en la tabla 4 la descripción de los tramos que se utilizarán en el desarrollo de este proyecto.

Tabla 4. Fechas de tramos temporales para COVID19

Tramo	Fecha Inicial	Fecha Final	Número de días
1	2020-03-02	2020-11-01	245
2	2020-11-02	2021-03-08	127
3	2021-03-09	2021-11-22	259
4	2021-11-23	2022-04-18	147

3.3.10. Análisis de descomposición por tramos.

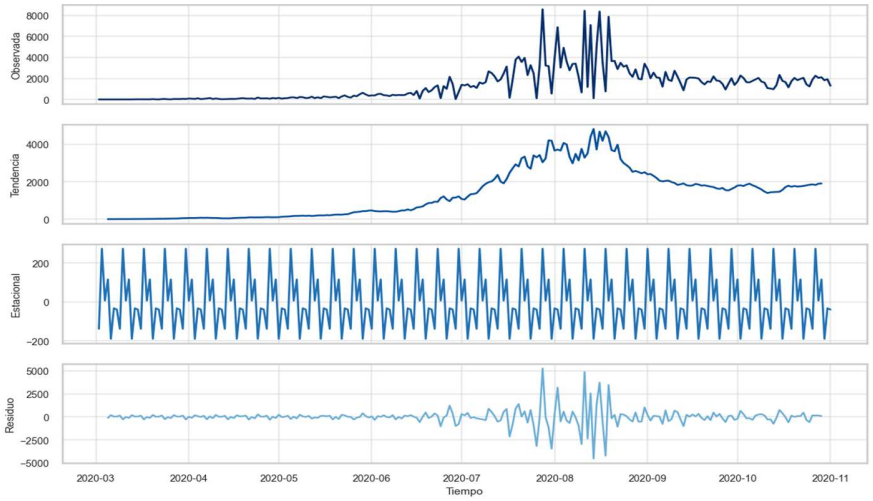
En la tabla 5 se muestra el análisis de estacionariedad de los cuatro tramos, seguido del análisis de descomposición de los mismos.

Tabla 5. Prueba de Dickey-Fuller Aumentada (ADF) y análisis de estacionariedad por tramos

Tramo	Estadístico ADF	p-valor	Estacionariedad (ADF)	Estacionalidad
Primer Tramo	-1.2357	0.6581	No estacionaria	Sí
Segundo Tramo	-1.2177	0.6659	No estacionaria	Sí
Tercer Tramo	-2.1228	0.2354	No estacionaria	Sí
Cuarto Tramo	-2.4775	0.1210	No estacionaria	Sí

En el primer tramo de la serie, los resultados de la prueba ADF indican notoriamente la ausencia de estacionariedad. El estadístico ADF obtenido es mayor, en valor absoluto, que los valores críticos para todos los niveles de significancia. Este resultado se ve reforzado por un valor p elevado, muy por encima del umbral convencional de 0.05. En consecuencia, la serie presenta una estructura temporal no estacionaria, característica común en procesos epidemiológicos durante fases de expansión o cambio dinámico. Los resultados se pueden apreciar en la figura 2.

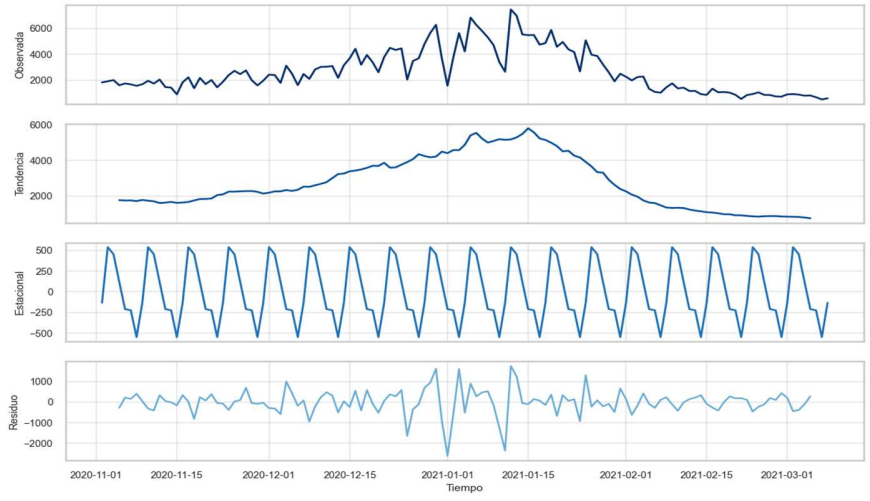
figura 2. Descomposición estacional del primer tramo



Este tramo es significativo en el sentido en que se empiezan a vislumbrar las características y las particularidades epidemiológicas del brote.

En el segundo tramo de la serie, la prueba de Dickey-Fuller aumentada evidencia nuevamente la no estacionariedad del proceso. El estadístico ADF no alcanza los valores críticos establecidos para todos los niveles de significancia, en este punto es claro que la media y la varianza de la serie no son constantes en el tiempo. Desde el punto de vista estructural, la evaluación automática de estacionalidad revela una amplitud estacional muy pronunciada. Este resultado sugiere la presencia de ciclos periódicos intensos, posiblemente asociados a dinámicas epidemiológicas recurrentes. En consecuencia, el tramo presenta una combinación de no estacionariedad y fuerte estacionalidad (figura 3).

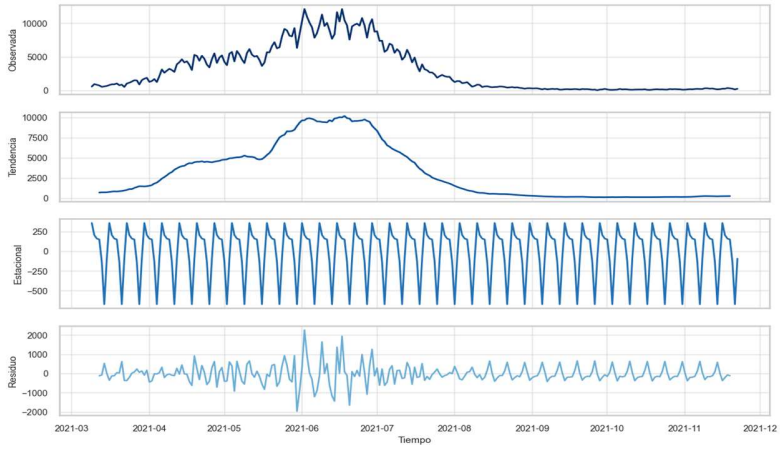
figura 3. Descomposición estacional del segundo tramo



En el tercer tramo (figura 4), la serie diaria de casos independientes de COVID19 en Bogotá, los resultados de la prueba ADF nuevamente apuntan a la falta de estacionariedad. El valor del estadístico

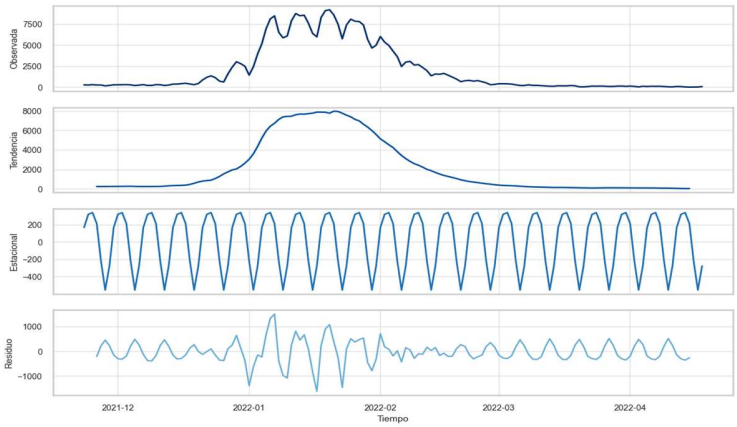
ADF se encuentra por encima de los valores críticos. Evidentemente, los resultados de la prueba refuerzan la conclusión de que la serie presenta características de no estacionariedad al mismo tiempo que indican la existencia de patrones estacionales significativos en la serie. Esta estacionalidad podría estar relacionada con la variabilidad en los casos de COVID19 debido a factores como las olas epidémicas o los ciclos de medidas de control.

figura 4. Descomposición estacional tercer tramo



Finalmente, en el cuarto tramo de la serie, los resultados de la prueba ADF indican que la serie no es estacionaria. El estadístico ADF no supera los valores críticos para ningún nivel de significancia, lo que significa que no se puede rechazar la hipótesis nula de la presencia de una raíz unitaria. La serie no exhibe propiedades de estacionariedad. Además, la evaluación automática de estacionalidad revela una amplitud estacional considerable muy superior al umbral mínimo (figura 5). Todo lo anterior confirma la existencia de patrones estacionales pronunciados en los datos.

figura 5. Descomposición estacional cuarto tramo



Los ciclos presentados en este tramo pueden estar relacionados con la propia dinámica del brote en un contexto de inmunización de la población avanzada. La presencia de estacionalidad, combinada con la no estacionariedad, implica que es necesario realizar transformaciones, diferenciaciones o modelos

estacionales, para garantizar que las predicciones sean precisas y capaces de capturar tanto la tendencia a largo plazo como los efectos cíclicos.

3.4. RESULTADO DE LA LIMPIEZA

Luego de realización de la depuración, el conjunto de datos quedó estructurado de forma robusta para su continuación y análisis. Este proceso permitió garantizar la integridad del conjunto de datos y su compatibilidad con técnicas avanzadas de modelado.

4. IDENTIFICACIÓN Y SELECCIÓN DE LAS VARIABLES CRÍTICAS

4.1. SELECCIÓN DE VARIABLES

La elección correcta de variables predictoras fue fundamental para mejorar el rendimiento de los modelos de predicción. Este capítulo se centra en identificar los atributos más influyentes del conjunto de datos sobre COVID19, evaluando su relevancia estadística y utilidad para la predicción de brotes.

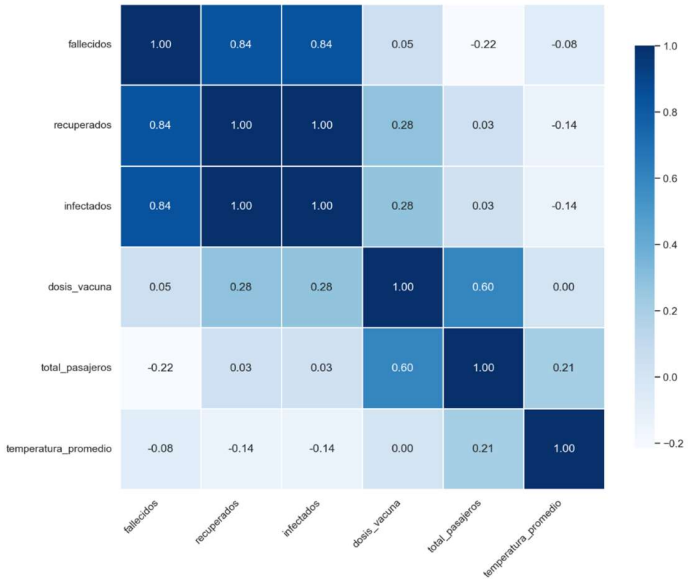
4.2. ESTRATEGIAS DE SELECCIÓN DE VARIABLES

- Análisis univariado y bivariado para descartar variables irrelevantes o altamente correlacionadas.
- Modelos de importancia de variables con XGBoost para obtener rankings objetivos.

4.3. ANÁLISIS DE CORRELACIÓN DE VARIABLES EXOGENAS

Para identificar la debilidad o fortaleza entre las variables, se aplicó el cálculo del coeficiente de correlación de Pearson, evidenciando una relación lineal débil entre la variable objetivo y las tres variables exógenas consideradas. Entre estas, la variable asociada al número de dosis de vacunas aplicadas presentó el mayor grado de correlación con la variable objetivo (0.28), seguida por la temperatura promedio de la ciudad (-0.14). En último lugar, se ubicó la variable correspondiente al total de pasajeros en tránsito por el Aeropuerto Internacional El Dorado de la ciudad de Bogotá (0.03). De la misma forma, no se presentaron correlaciones significativas entre estas covariables durante el período de análisis (figura 6).

figura 6. Matriz de correlación de las variables



Dada la baja magnitud de las correlaciones observadas entre las variables exógenas y la variable a predecir, fue necesario recurrir a la construcción de nuevas variables para mejorar el ajuste de los modelos. No obstante, aunque las correlaciones lineales no resultaron significativas, estas proporcionaron criterios orientadores para la selección de variables relevantes en el proceso de modelado. En consecuencia, dichas variables fueron incorporadas en la construcción de características con el objetivo de identificar posibles relaciones no lineales, particularmente en aquellas asociadas a efectos de rezago que no son evidentes a simple vista.

4.4. INGENIERÍA DE CARACTERÍSTICAS

En vista de la débil correlación observada en la figura 6 entre las variables exógenas originales, se procedió a la generación de 247 nuevas variables mediante técnicas de ingeniería de características. Estas variables fueron diseñadas con el propósito de capturar dependencias temporales, dinámicas no lineales y patrones latentes que no resultaban evidentes en el conjunto de variables iniciales.

El objetivo principal de esta ampliación del espacio de características fue entrenar modelos preliminares que permitieran establecer "rankings" objetivos basados en la importancia relativa de las variables en el proceso de ajuste. En consecuencia, dichos rankings facilitaron una selección informada y sistemática de las variables más relevantes, contribuyendo a la optimización del desempeño predictivo y la estabilidad de los modelos finales (Tabla 6).

Tabla 6. Características de las variables creadas.

Tipo de Variable	Número Aproximado	Características
Original	1+ (variable objetivo y otras numéricas)	Datos originales del DataFrame, sin transformación.
Lags de la variable objetivo	3 (según lags: 1, 7, 14)	Valores desplazados hacia atrás para capturar dependencia temporal.
Diferencias de lags	3	Diferencias entre el valor actual y el valor rezagado.
Cambio porcentual de lags	3	Incremento relativo respecto al valor rezagado.
Rolling statistics (media, std, max, min)	8 (2 ventanas × 4 estadísticas)	Promedios, desviaciones, máximo y mínimo móvil para capturar tendencia local.
Z-score rolling	2 (1 por cada ventana)	Normalización centrada en media y desviación móvil.
Crecimiento porcentual diario	1	Variación porcentual día a día de la variable objetivo.

Aceleración (segunda diferencia)	1	Segunda diferencia temporal, capturando cambios de pendiente.
Cambio porcentual rolling	2 (1 por cada ventana)	Cambio relativo respecto al valor de la ventana previa.
Ratio respecto al máximo rolling	2 (1 por cada ventana)	Relación de valor actual con máximo de la ventana.
Calendario: día de la semana	1	Valor de 0 a 6, representando el día de la semana.
Calendario: fin de semana	1	Indicador binario: 1 si es sábado o domingo, 0 si no.
Calendario: mes	1	Valor de 1 a 12 del mes correspondiente.
Multivariantes: lags de otras variables	3×n_vars(excluyendo objetivo y columnas excluidas)	Lags de variables numéricas adicionales.
Multivariantes: rolling stats de otras variables	2×3× n_vars = 4× n_vars	Rolling mean y std para cada ventana y variable numérica adicional.

5. ENTRENAMIENTO Y ESTIMACIÓN DE LOS MODELOS

El presente capítulo describe el proceso de entrenamiento y estimación de los modelos predictivos desarrollados en este estudio, abordando de manera sistemática los fundamentos teóricos y operativos que sustentan su construcción. En cada sección se detallan las ecuaciones que definen cada enfoque, los parámetros involucrados, las métricas utilizadas para evaluar su desempeño y el procedimiento seguido para la calibración y validación de los modelos. Asimismo, se presenta el paso a paso de la obtención de los modelos, desde el preprocesamiento de los datos y la selección de variables relevantes, hasta la estimación de los parámetros y la evaluación de la capacidad predictiva. Este enfoque metodológico permite garantizar la reproducibilidad de los resultados y facilita la comparación objetiva entre los modelos matemáticos, estadísticos y de aprendizaje automático, proporcionando un marco riguroso para el análisis de la dinámica del COVID19 en la ciudad de Bogotá.

5.1. MODELO MATEMÁTICO

Para el modelo SEIR, se inició con el ajuste de los parámetros que el sistema conlleva. Inicialmente se definieron los parámetros poblacionales y de vacunación $(N, \mu\ y\ \nu)$, que son parámetros estructurales del modelo SEIR que no dependen de los datos de casos, sino de características de la ciudad, que para este fin es Bogotá y de la campaña de vacunación.

Se estima una población total susceptible al inicio de $N = 7'900.000$ habitantes, relacionado con las proyecciones oficiales del DANE, una tasa demográfica μ que representa la tasa natural de entrada y salida de individuos del sistema (nacimientos y muertes no relacionadas con la enfermedad). Para este proceso se asume una esperanza de vida promedio de 75 años, por lo tanto:

$$\mu \approx \frac{1}{75 * 365} \approx 3.65\ x\ 10^{-5}\ por\ dia \tag{18}$$

La tasa de vacunación ν fija derivada de los datos diarios de vacunación. La serie es convertida bajo la serie real de vacunación en un parámetro ν constante para el modelo. Para ello, se definió la eficacia vacunal efectiva (η) que representa el porcentaje de vacunados bajo la premisa de que realmente adquieren inmunidad efectiva, fijándose este valor en un 80%, valor razonable para esquemas iniciales de COVID19; seguido se determinó una vacunación promedio diaria calculando el promedio de toda la campaña de vacunación, para lo cual se asumió que el ritmo de vacunación podría ser modelado como un flujo constante en lugar de una serie variable día a día, para finalmente estipular la tasa efectiva de vacunación:

$$\nu = \eta * \frac{Vacunación\ promedio\ diaria}{N} \tag{19}$$

Donde ν es la probabilidad de que un individuo susceptible pase al estado inmune por vacunación. El ajuste de los parámetros β, σ, γ que son los que relaciona el modelo y ρ , $E0_{frac}\ y\ I0_{frac}$ son parámetros adicionales para el funcionamiento de este, donde ρ es la proporción de infecciones reales que se

reportan como casos observados (tasa de reporte), $E0_{frac}$ es la proporción inicial de personas infectadas pero aún no infecciosas (expuestos) y $E0_{frac}$ es la proporción inicial de personas infecciosas el primer día del modelo, se definieron mediante una función objetivo que calibró el modelo SEIR con vacunación constante en toda la serie temporal diaria. La función recibe un conjunto de parámetros tentativos, construye el modelo correspondiente y lo simula para reproducir los datos diarios de Bogotá. Luego de esto, se calcula la diferencia entre la incidencia modelada y la observada mediante un error logarítmico, adecuado para datos ruidosos, con ceros y variabilidad extrema. Estos residuos alimentan el algoritmo de mínimos cuadrados no lineales (*least_squares)*, que ajusta los parámetros $\beta, \sigma, \gamma, \rho, E0$ e $I0$ para lograr el mejor ajuste posible del modelo a los datos reales.

En el proceso de calibración del modelo SEIR, se estimó el estado inicial del sistema el primer día del periodo analizado S_0 , E_0 , I_0 , R_0 sin embargo como no se tienen datos reales directos que indiquen cuántos expuestos (E_0) o infectados activos (I_0) había exactamente en Bogotá al inicio de la epidemia, por lo tanto, se estimaron como parámetros del modelo:

$$\begin{gathered} E_0 = E0_{frac}.N \\ I_0 = I0_{frac}.N \\ R_0 = 0 \\ S_0 = N - (E_0 + I_0 + R_0) \end{gathered} \tag{20}$$

Para este proceso, se asume que $R_0 = 0$ ya que al inicio de la pandemia no se registran personas recuperadas.

Los valores iniciales para los parámetros fueron definidos en base a la literatura epidemiológica ampliamente aceptada. El período de incubación se fijó en 5.2 días (26), mientras que la duración infecciosa promedio se asumió de 5 días (27). La tasa de transmisión inicial se estableció en un valor moderado ($\beta = 0.3$), coherente con estudios SEIR anteriores (28). Para el parámetro de reporte se utilizó $\rho = 0.2$, consistente con estimaciones de subregistro que indican que hasta el 80% de las infecciones no fueron documentadas (29). Finalmente, las condiciones iniciales se definieron como fracciones pequeñas de la población, siguiendo prácticas estándar en modelación epidemiológica (30), Por tanto, los parámetros determinados son:

- $\beta = 0.3$
- $\sigma = {}^{1}/_{5.2}$
- $\gamma = {}^{1}/_{5}$
- $\rho = 0.20$
- $E0_{frac} = 5\ x\ 10^{-6}$
- $I0_{frac} = 3\ x\ 10^{-6}$

De igual manera, estos parámetros se imponen sobre unas restricciones epidemiológicamente razonables:

- $\beta\ entre\ 0.05\ y\ 1.5$

- $\sigma\ entre\ {}^{1}/_{14}\ (incubación\ hasta\ 14\ dias)\ y\ {}^{1}/_{3}\ (incubación\ mínima\ de\ 3\ dias)$
- $\gamma\ entre\ {}^{1}/_{14}\ (infeccioso\ hasta\ 14\ dias)\ y\ {}^{1}/_{2}\ (duración\ mínima\ de\ 2\ dias)$
- $\rho\ entre\ 0.001\ y\ 1.0\ (entre\ 0.1\%\ y\ 100\%\ de\ infecciones\ reportadas)$
- $E0_{frac}$ y $I0_{frac}\ entre\ 1\ x\ 10^{-8}\ y\ 1\ x\ 10^{-2}\ (fracciones\ muy\ pequeñas)$

Esto evita que el optimizador encuentre soluciones absurdas (por ejemplo, β gigantes o σ negativas) asegurando que los parámetros permanezcan en rangos biológicamente plausibles.

Los parámetros epidemiológicos estimados para la serie completa están en la tabla 7.

Tabla 7. Parámetros iniciales modelo continuo

β	σ	γ	ρ	v	μ	E_0	I_0
0.20	0.07	0.07	0.99	0.0019	0.000037	0.18	0.65

Evaluando el modelo calibrado, arrojó el siguiente comportamiento (figura 7) y determino las siguientes métricas (Tabla 8):

figura 7. Gráfica SEIR continuo

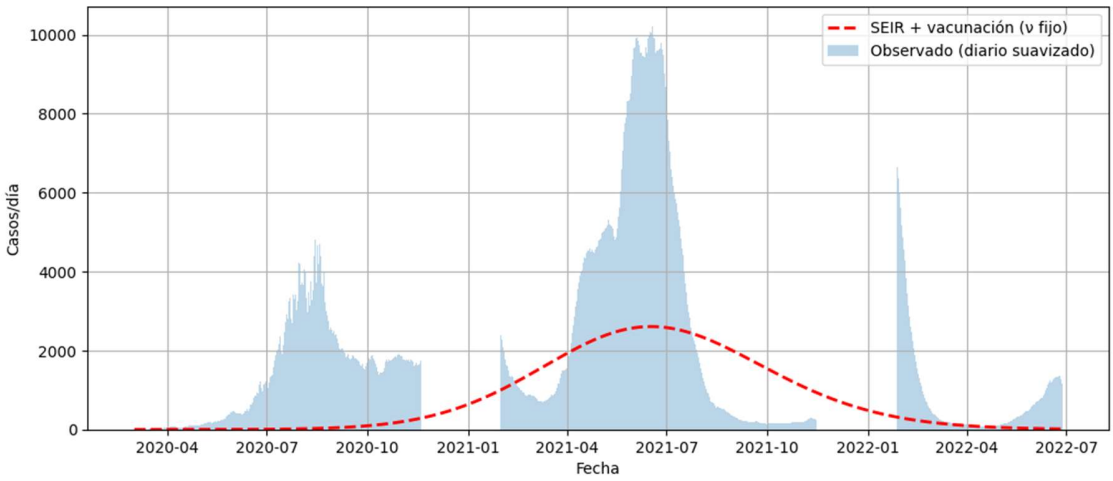


Tabla 8. Métricas modelo continuo

MAE	$RMSE$	R^2	R_0
1769.6	2590.3	−0.093	2.78

La figura 7 muestra un comportamiento claro, el modelo SEIR continuo genera una sola ola grande y suave, mientras que los datos reales presentan cuatro olas bien diferenciadas, con aumentos y caídas abruptas. La realidad de Bogotá durante el COVID19 no fue homogénea, se presentaron múltiples cambios abruptos en transmisibilidad, comportamiento social, vacunación y aparición de nuevas

variantes. Dado que el modelo asume valores constantes para β y ν, la dinámica resultante solo puede producir una ola epidémica suave, sin posibilidad de generar picos sucesivos. Por esta razón, el ajuste continuo no reproduce las cuatro olas observadas y la curva modelada aparece suavizada y desfasada respecto a los datos reales.

El número reproductivo básico estimado indica que, en promedio, cada persona infectada habría contagiado aproximadamente a tres personas en una población completamente susceptible, esto presenta un nivel de transmisibilidad moderado alto, consistente con los valores reportados en la literatura para SARS-CoV-2 durante la primera etapa de la pandemia.

Como ya se indicó anteriormente, dado que la serie de casos para Bogotá presenta múltiples olas epidémicas, se vuelve necesario ajustar el modelo de manera segmentada por tramos, donde cada tramo refleja cambios sustanciales en las condiciones de transmisión que modifican los parámetros epidemiológicos del modelo SEIR, por tanto, un único conjunto de parámetros no puede describir adecuadamente toda la pandemia. Al ajustar cada tramo por separado, cada una captura la dinámica específica de ese periodo, produciendo métricas de ajuste significativamente superiores y resultados epidemiológicamente coherentes.

Se realizó el mismo procedimiento que con el modelo continuo, se ajustó el modelo solo dentro de la ventana temporal del tramo, permitiendo que cada una tenga sus propios parámetros $\beta, \sigma, \gamma, \rho, E0_{frac}$, $I0_{frac}$ y una tasa de vacunación ν fija específica de ese tramo, calculada a partir del promedio de vacunación diaria. Una vez definidas las funciones de ajuste para cada ola, se procedió a ejecutar el modelo de forma independiente en cada uno de los cuatro tramos epidémicos seleccionados. Para ello se generan las métricas de desempeño y la gráfica correspondiente.

5.1.1. MODELO PRIMER TRAMO

En el primer tramo se ajustaron los parámetros (tabla 9) correspondientes con las fechas establecida (tabla 4).

Tabla 9. Parámetros primer tramo

$\boldsymbol{\beta}$	$\boldsymbol{\sigma}$	$\boldsymbol{\gamma}$	$\boldsymbol{\rho}$	$\boldsymbol{\nu}$	$\boldsymbol{\mu}$	$\boldsymbol{E_0}$	$\boldsymbol{I_0}$
0.19	0.07	0.07	0.04	0.0000	0.000037	0.32	2400.87

En este tramo, un $\beta \approx 0.19$ indica un contagio moderado, compatible con restricciones iniciales y baja movilidad, un $\sigma \approx \gamma \approx 0.07$ equivalente a una incubación y una duración infecciosa de 14 días, un $\rho \approx 0.04$ ya que en el año 2020 existían pocos registros, un $\nu \approx 0$ indica la no presencia de vacunas y un $E0$ y $I0$ pequeños, indican epidemia en fase inicial. A partir de allí se ejecutó el modelo para el primer tramo (figura 8) con sus métricas (tabla 10).

figura 8. Gráfica primer tramo

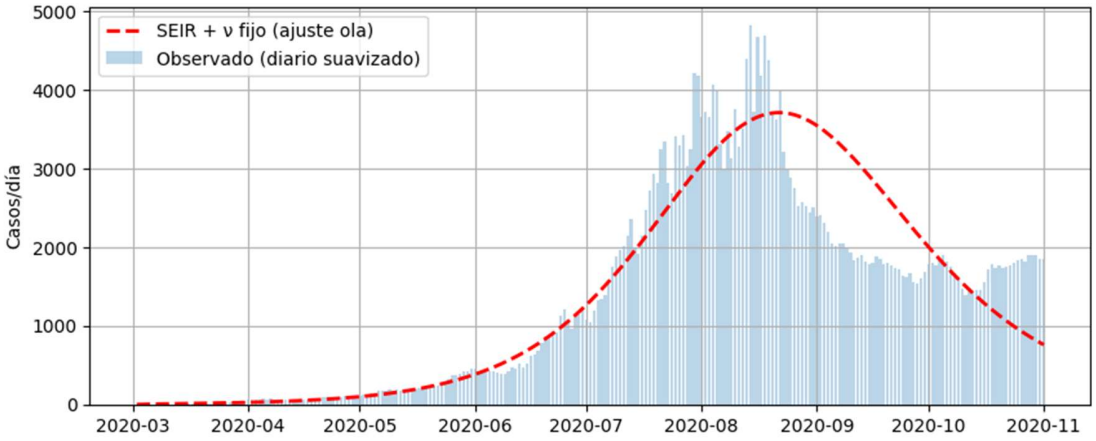


Tabla 10. Métricas primer tramo

MAE	$RMSE$	R^2	R_0
322.4	523.5	0.82	2.65

El modelado arrojó un $R^2 \approx 0.82$, entendiéndose como un buen ajuste, pues el primer tramo fue el más "clásico" y el modelo SEIR pudo representarlo de la mejor manera, un $R_0 \approx 2.65$ de alta transmisibilidad indicando un sistema no inmune. Sin embargo, cuando se evaluó la capacidad predictiva del modelo SEIR para dicho tramo no es buena. Cabe resaltar que a diferencia de los modelos anteriores donde se consolida un 80% de entrenamiento y 20% prueba, acá para el modelo SEIR se definió un horizonte de predicción del 20% (en días) por tramo y, a partir de los parámetros calibrados, se simuló la dinámica del sistema hasta el final del tramo y posteriormente se proyectaron varios días hacia adelante (figura 9). Esta gráfica muestra visualmente que los casos observados (tramo y post-tramo) y las curvas simuladas de ajuste y predicción se alejan, lo que permite evaluar visualmente el modelo. Además, se calculan las métricas de predicción (tabla 11) comparando las proyecciones del modelo con los datos reales posteriores.

figura 9. Predicción primer tramo

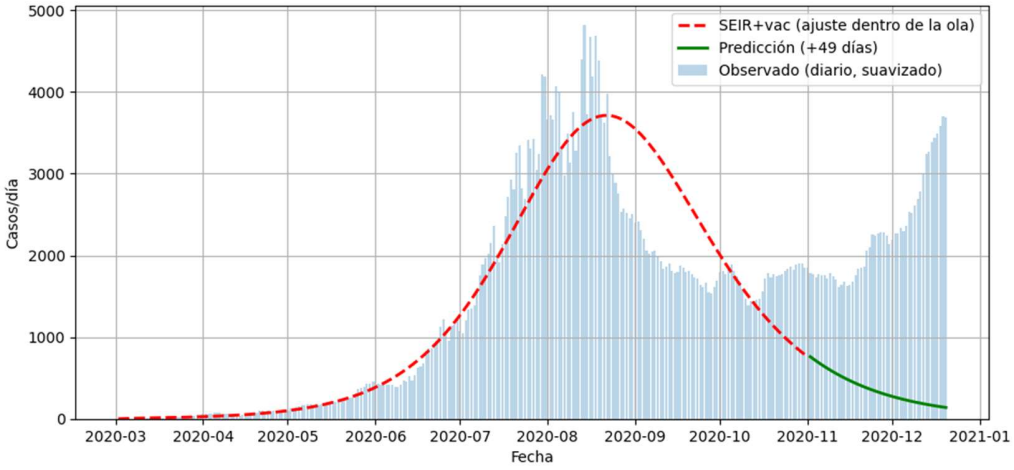


Tabla 11. Métricas de predicción primer tramo

MAE	$RMSE$	R^2
$\mathbf{1922.6}$	2073.6	-10.11

Se identifica un $R^2 < 0$, lo que significa una predicción muy mala para el primer tramo.

5.1.2. MODELO SEGUNDO TRAMO

En el segundo tramo se ajustaron los parámetros (tabla 12) para el modelo SEIR.

Tabla 12. Parámetros segundo tramo

β	σ	γ	ρ	v	μ	E_0	I_0
$\mathbf{0.30}$	0.33	0.21	0.08	0.000062	0.000037	79000	0.74

En este tramo arroja un $\beta \approx 0.30$ indicando un incremento claro en la transmisibilidad, acorde con flexibilización social a finales del año 2020 y festividades, un $\sigma \approx 0.33$ equivalente a una incubación de 3 días, un $\gamma \approx 0.21$ que indica una duración infecciosa de 5 días, un $\rho \approx 0.08$ de mejor capacidad diagnóstica, un $v \approx 0.00062$ inicio de vacunación y un $E0$ y $I0$ ya en aumento. Con esto se ejecutó el modelo (figura 10) y las métricas (tabla 13) para el segundo tramo.

figura 10. Gráfica segundo tramo

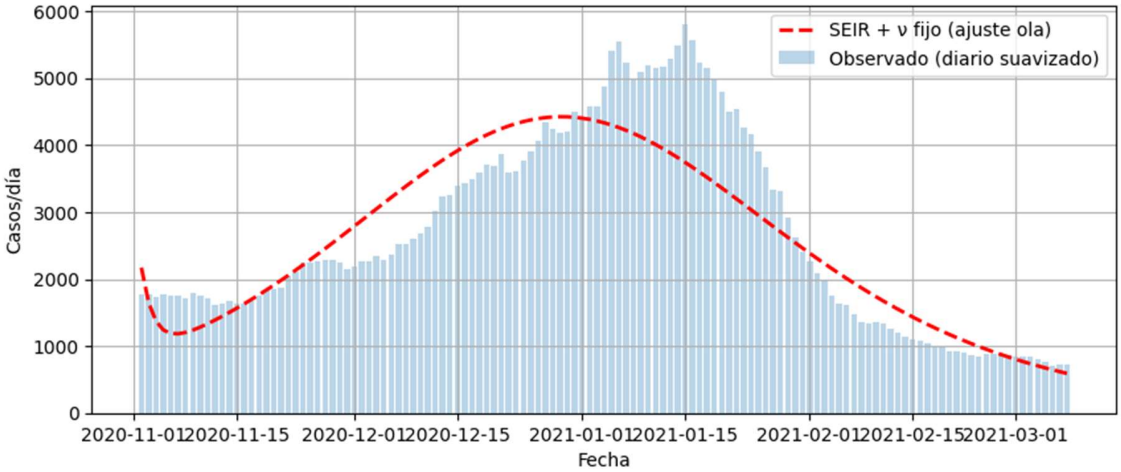


Tabla 13. Métricas segundo tramo

MAE	$RMSE$	R^2	R_0
$\mathbf{501.5}$	672.7	0.79	1.39

El modelado en el segundo tramo arrojó un $R^2 \approx 0.79$, que es un ajuste bueno; con una dinámica más estable. El $R_0 \approx 1.39$ se interpreta como disminución por medidas y comportamiento social. Evaluando la capacidad predictiva del modelo SEIR para este tramo (figura 11) y sus métricas (tabla 14), igualmente se observa un $R^2 < 0$ entendiéndose como una predicción mala.

figura 11. Predicción segundo tramo

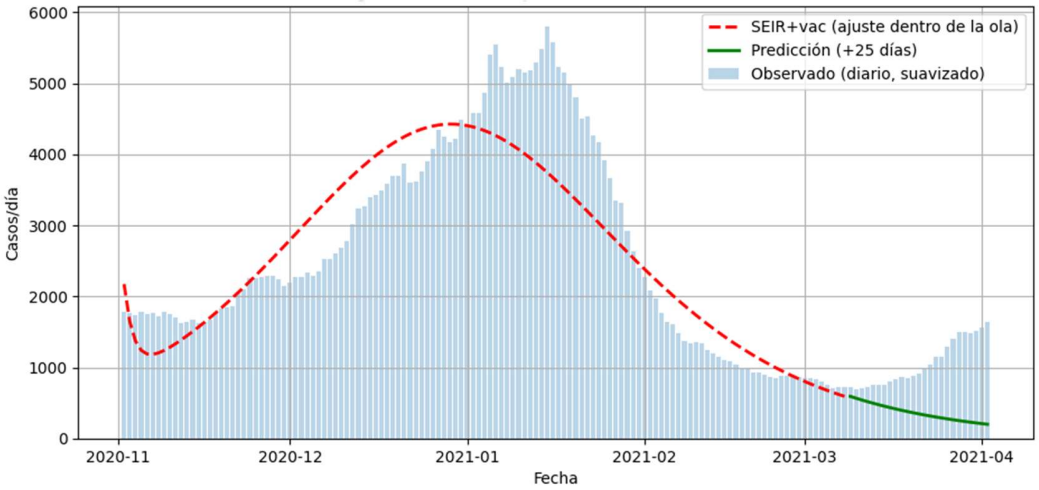


Tabla 14. Métricas de predicción segundo tramo

MAE	$RMSE$	R^2
$\mathbf{696.8}$	820	-5.50

5.1.3. MODELO TERCER TRAMO

Para el tercer tramo se ajustaron los parámetros correspondientes (tabla 15), obteniendo un $\beta \approx 0.18$ donde nuevamente disminuye; este tramo se puede asociar a cambios en movilidad y variantes, pero el modelo captura el menor β de todos los tramos, un $\sigma \approx \gamma \approx 0.07$ equivalente a una incubación y una duración infecciosa de 14 días; un $\rho \approx 0.14$ de aumento de pruebas, un $v \approx 0.0037$ intensificación en la vacunación y un $E0\ y\ I0$ aumentando. Se ejecuta el modelo para el tercer tramo (figura 12) y sus métricas correspondientes (tabla 16).

Tabla 15. Parámetros tercer tramo

β	σ	γ	ρ	v	μ	E_0	I_0
$\mathbf{0.18}$	0.07	0.07	0.14	0.0037	0.000037	56139.6	79000

figura 12. Gráfica tercer tramo

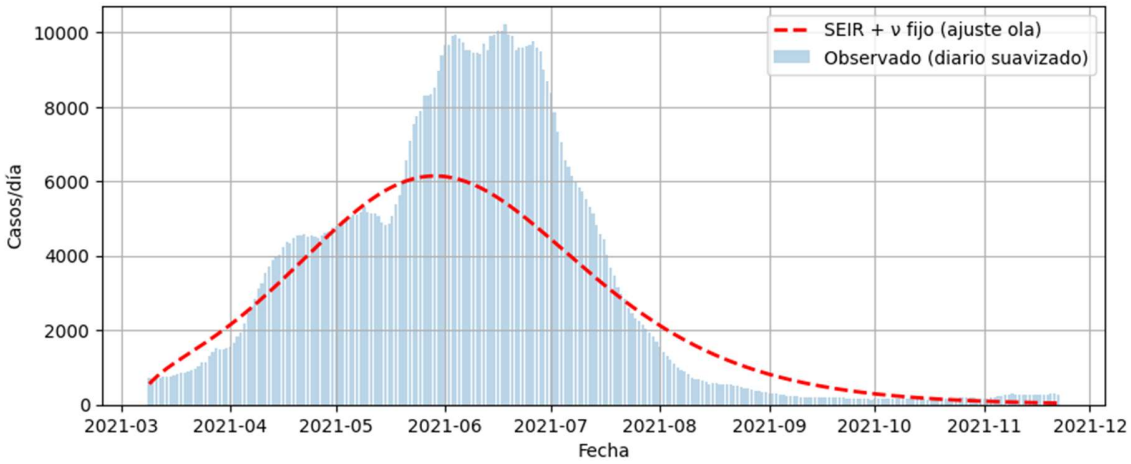


Tabla 16. Métricas tercer tramo

MAE	$RMSE$	R^2	R_0
980.5	1650	0.75	2.59

El modelo en el tercer tramo arrojó un $R^2 \approx 0.75$, indicando un buen ajuste, pero menor precisión por duración prolongada. El $R_0 \approx 2.59$ reflejando presencia probable de variantes más contagiosas. Ajustando la capacidad predictiva del modelo SEIR para este tramo (figura 13) y las métricas para esta predicción (tabla 17) se observa que $R^2 < 0$ entendiéndose como una predicción igualmente deficiente a los tramos anteriormente analizados.

figura 13. Predicción tercer tramo

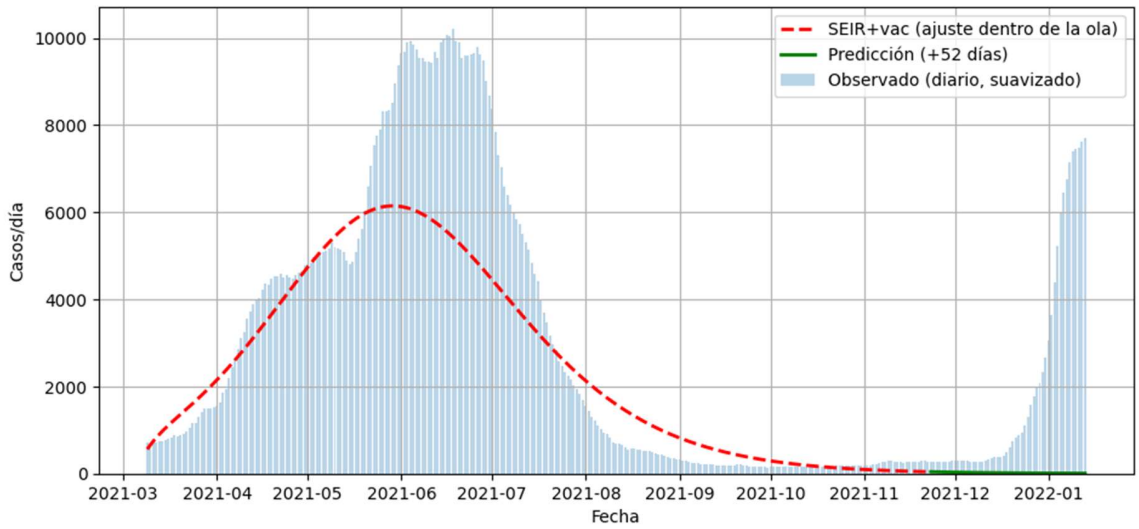


Tabla 17. Métricas de predicción tercer tramo

MAE	$RMSE$	R^2
$\mathbf{2045.6}$	3284.8	-0.64

5.1.4. MODELO CUARTO TRAMO

En el último tramo se ajustaron los parámetros (tabla 18) encontrando un $\beta \approx 0.48$ que implica un máximo contagio estimado en este periodo. Este tramo coincide con variantes más transmisibles como Ómicron, un $\sigma \approx \gamma \approx 0.07$ equivalente a una incubación y una duración infecciosa de 14 días, un $\rho \approx 0.04$ de disminución de pruebas, un $v \approx 0.0034$ intensificación los programas de vacunación y un $E0$ y $I0$ en aumento.

Tabla 18. Parámetros cuarto tramo

β	σ	γ	ρ	v	μ	E_0	I_0
$\mathbf{0.48}$	0.07	0.07	0.04	0.0034	0.000037	55261	0.26

Se ejecuta el modelo para el cuarto tramo (figura 14) y sus métricas (tabla 19). El modelo arroja un $R^2 \approx 0.85$, definiendo un ajuste bueno pese a la velocidad extrema de Ómicron. El $R_0 \approx 6.67$ interpretándose como un valor extremadamente alto, lo cual coincide con Ómicron.

figura 14. Gráfica cuarto tramo

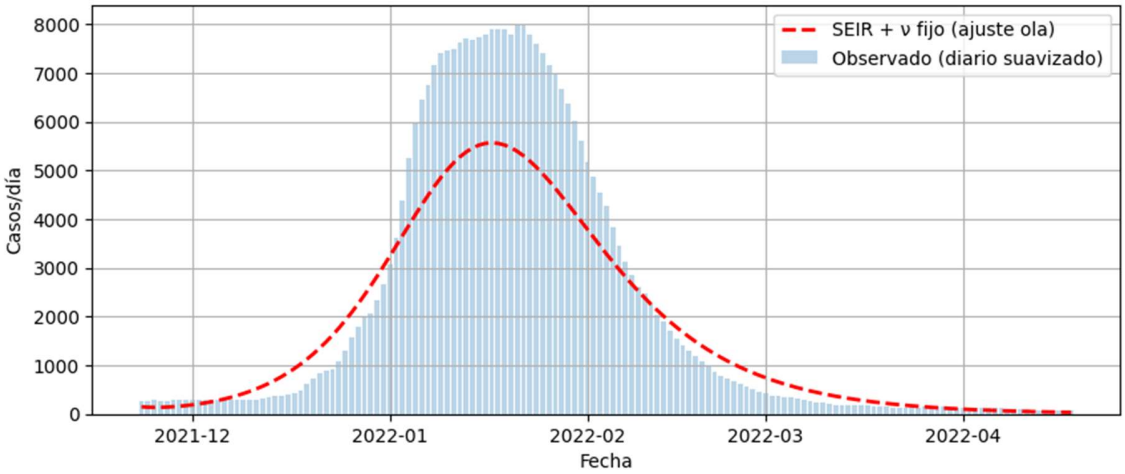


Tabla 19. Métricas cuarto tramo

MAE	$RMSE$	R^2	R_0
$\mathbf{631.3}$	1048.9	0.85	6.67

Ahora, ajustando la capacidad predictiva del modelo SEIR para este último tramo (figura 15) y las métricas de predicción (tabla 20) se tiene igualmente como en los tramos anteriores, un $R^2 < 0$ entendiéndose como una predicción bastante deficiente.

figura 15. Predicción cuarto tramo

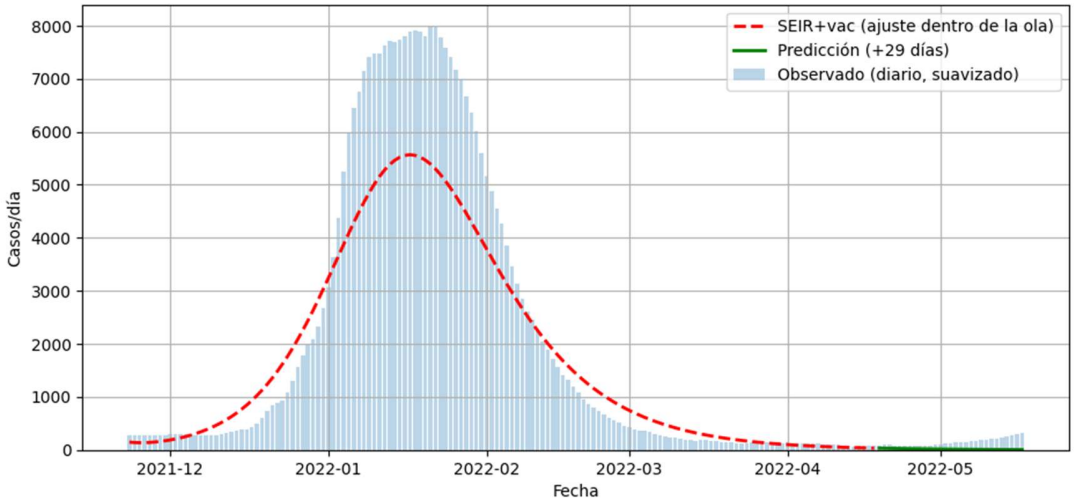


Tabla 20. Métricas de predicción cuarto tramo

MAE	$RMSE$	R^2
136.2	156.9	-3.862

En la tabla 21 se muestra un resumen de los parámetros determinados, destacando contagios moderados (β) tanto en el primer y tercer tramo compatible con la baja movilidad y restricciones, mientras que en el segundo y cuarto tramo hubo un incremento claro en la trasmisibilidad, indicando el cuarto tramo como el de máximo contagio estimado en el periodo estudiado.

Tabla 21. Resumen de parámetros y métricas

Tramo	β	σ	γ	ρ	μ	v	R_0	MAE	RMSE	R^2
1	0.19	0.07	0.07	0.04	0.000037	0.000000	2.65	322.42	523.49	0.82
2	0.30	0.33	0.21	0.08	0.000037	0.000062	1.39	501.52	672.65	0.79
3	0.18	0.07	0.07	0.14	0.000037	0.0037	2.59	980.51	1650	0.75
4	0.48	0.07	0.07	0.04	0.000037	0.0034	6.67	631.31	1048.93	0.85

El último tramo coincide con las variantes más transmisibles como Ómicron, es decir, coincide con la evidencia epidemiológica real. En cuanto a σ (velocidad de avance del estado expuesto) tiene valores entre 0.07 y 0.32, que corresponde a periodos de incubación de 14 días y 3 días, respectivamente, lo que indica que son valores epidemiológicamente plausibles, pues variantes posteriores como Delta y Ómicron mostraron períodos de incubación más cortos, lo cual coincide con el segundo tramo (σ más alto). Con γ (recuperación) varía entre 0.07 y 0.21, que equivale aproximadamente a 14 y 5 días, esto muestra la dificultad del modelo para ajustar olas con dinámicas rápidas, obligando a modificar la duración infecciosa para compensar efectos estructurales del sistema.

Con respecto a ρ(fracción de casos observados), oscila entre 0.03 y 0.14, pues en el primer y cuarto tramo hubo un subregistro mayor, y en el segundo y tercer tramo hubo una mejor capacidad diagnóstica y aumento de pruebas. Esta tendencia también coincide con la historia real: en 2020 había poco testeo;

en 2021 la disponibilidad de pruebas aumentó.

En cuanto a la tasa de vacunación efectiva v, los valores corresponden perfectamente al periodo real: el primer tramo no había vacunas, el segundo tramo inicio de la vacunación (tasa baja) y el tercer y cuarto tramo intensificaron el programa de vacunación.

El R_0 básico por tramo: el primer y tercer tramo tienen una alta transmisibilidad; refleja presencia probable de variantes contagiosas; el segundo tramo, una disminución por medidas y comportamiento social; y el cuarto tramo, extremadamente alto, coincide con Ómicron.

Por último, con respecto a las métricas, el primer y cuarto tramo tuvieron un buen ajuste. El primer tramo el modelo SEIR lo pudo representar bien. El cuarto tramo se ajustó bien pese a la velocidad extrema de la variante. En el segundo y tercer tramo, el ajuste fue bueno, las dinámicas fueron más estable, sin embargo, las predicciones en todos los tramos arrojan resultados desfavorables.

Además de las métricas tradicionales (MAE, RMSE y R^2), se emplearon las métricas MAPE y MASE (tabla 22) para evaluar el desempeño predictivo del modelo.

Tabla 22. Métricas MAPE y MASE

Tramo	MAPE	MASE
1	91.00	3.28
2	18.23	5.03
3	55.66	10.83
4	41.99	5.75

En términos de MAPE, se observa un desempeño heterogéneo entre los cuatro tramos. El segundo tramo presenta el mejor resultado (MAPE ≈ 18 %), lo que indica que, durante este periodo, el modelo logró una aproximación razonable a los valores observados en términos relativos. En contraste, el primer, tercer y cuarto tramo exhiben valores elevados de MAPE (entre 42 % y 91 %), asociados principalmente a la presencia de valores bajos de incidencia al inicio o al final de los tramos, así como a una mayor volatilidad en los casos diarios.

Por su parte, la métrica MASE muestra valores mayores que 1 en todos los tramos, lo que implica que, a escala diaria, el modelo SEIR no supera sistemáticamente a un modelo de referencia ingenuo basado en persistencia temporal. Este resultado es especialmente marcado en el tercer tramo, donde el MASE alcanza su valor máximo (≈ 10.83), reflejando la dificultad del modelo para capturar dinámicas complejas asociadas a cambios estructurales en la transmisión, tales como la introducción de nuevas variantes y el avance acelerado del proceso de vacunación. Incluso en los tramos con mejor ajuste visual y altos valores de R^2, el MASE indica que la mejora predictiva diaria frente a un modelo simple es limitada.

El ajuste diario por tramos muestra que el modelo SEIR con vacunación constante (v) reproduce adecuadamente la forma interna de cada tramo, obteniendo valores de R^2 entre $0.75\ y\ 0.85$. Esto indica que, cuando se analizan tramos epidemiológicamente homogéneos, el modelo capta satisfactoriamente

la dinámica local de transmisión. Sin embargo, las predicciones hacia adelante presentan valores de R^2 negativos, lo que refleja una capacidad predictiva limitada, lo que implica que, entre tramos, ocurren cambios sustanciales en el comportamiento social, aparición de nuevas variantes, modificaciones en las tasas reales de vacunación y alteraciones en la movilidad; todos ellos fenómenos que el modelo no incorpora explícitamente y que modifican drásticamente los parámetros epidemiológicos. Por tanto, aunque el modelo es útil para describir y ajustar cada ola de forma independiente, no es adecuado para realizar predicciones confiables entre tramos sin considerar un esquema más complejo con $\beta(t)$, variantes, o un componente estocástico o multicompartimental más realista.

5.2. MODELOS ESTADÍSTICOS.

Los modelos estadísticos constituyen una piedra angular en el análisis y predicción de fenómenos epidemiológicos, ofreciendo un marco metodológico robusto para comprender la dinámica de transmisión de enfermedades infecciosas. A diferencia de los modelos matemáticos los cuales se basan en supuestos teóricos de mecanismos de transmisión, los modelos estadísticos se fundamentan en patrones empíricos extraídos de datos observados, permitiendo capturar relaciones complejas entre variables sin requerir un conocimiento exhaustivo de los mecanismos subyacentes del proceso infeccioso (13).

5.2.1. MODELOS PRIMER TRAMO

En el contexto de la pandemia por COVID19, la modelación estadística ha demostrado para el contexto urbano como Bogotá, se desarrolló un modelo de pronóstico para la serie temporal de infectados diarios por COVID19, utilizando metodologías ARIMA/SARIMA. El modelo final seleccionado captura efectivamente la estacionalidad semanal y los patrones autoregresivos de la serie, demostrando alta capacidad predictiva.

Ecuación estimada de un modelo ARIMA(1,1,2)

El modelo se ajusta sobre la serie diferenciada. La ecuación del modelo es:

$$\Delta y_t = \phi_1 \Delta y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \quad (21)$$

donde:

- $\Delta y_t = y_t - y_{t-1}$ representa el cambio diario en la variable de interés
- ϕ_1 es el coeficiente autorregresivo de primer orden
- θ_1 y θ_2 son los coeficientes de media móvil
- ε_t término de error aleatorio con media cero y varianza constante

Forma equivalente expandida:

$$y_t - y_{t-1} = \phi_1 (y_{t-1} - y_{t-2}) + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \quad (22)$$

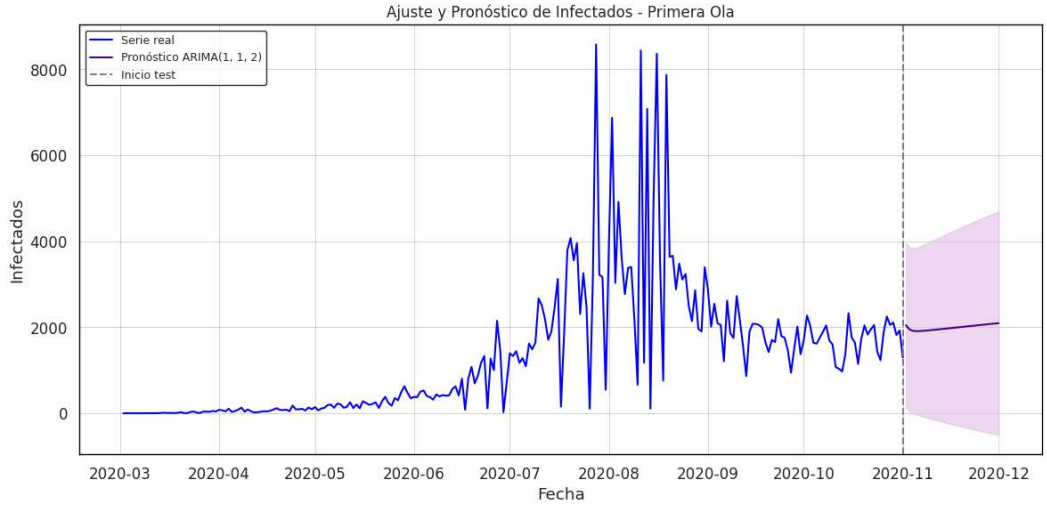
donde la variable dependiente corresponde a la primera diferencia de la serie original, y los términos autorregresivos y de media móvil capturan la dependencia temporal de corto plazo observada en los datos.

El modelo ARIMA(1,1,2) fue el mejor después de 8 combinaciones y corresponde a un modelo autorregresivo integrado de media móvil con un término autorregresivo de orden 1 indica que el cambio diario en los contagios depende linealmente del cambio observado en el día inmediatamente anterior. La inclusión de este término fue sugerida por la función de autocorrelación parcial (PACF), que mostró un corte significativo en el primer rezago; una diferenciación de primer orden que se aplicó para eliminar la tendencia presente en la serie original y garantizar estacionariedad en la media. Este valor fue determinado mediante pruebas de raíz unitaria (ADF) y verificación visual de la serie diferenciada y dos términos de media móvil, MA(1) y MA(2) los cuales capturan la dependencia del proceso con errores aleatorios de uno y dos días previos, respectivamente.

Análisis del ajuste y comportamiento predictivo del modelo ARIMA(1,1,2)

La figura 16 presenta el ajuste del modelo ARIMA(1,1,2) y muestra una alta coherencia entre la serie observada de nuevos casos diarios y la serie ajustada por el modelo, especialmente durante las fases de crecimiento y descenso de la ola epidémica. El modelo logra capturar adecuadamente la tendencia ascendente inicial, así como el cambio de régimen posterior al pico, lo que evidencia una correcta especificación del orden de diferenciación y de los componentes autorregresivos y de media móvil.

figura 16. Predicción primer tramo



Durante el periodo del pico, se observa que el modelo suaviza los valores extremos de la serie real. Este comportamiento es consistente con la naturaleza de los modelos ARIMA, los cuales priorizan la estructura promedio de la serie sobre la reproducción exacta de picos abruptos. Desde una perspectiva predictiva, esta suavización reduce el riesgo de sobreestimación extrema, aunque limita la precisión en la anticipación de máximos diarios.

En términos visuales, al representarse en tonos azules, la comparación entre la serie real y la serie ajustada resalta con mayor claridad la cercanía del ajuste, evitando contrastes agresivos y permitiendo una lectura más limpia del comportamiento temporal. El uso de una paleta monocromática facilita la interpretación del modelo como una aproximación estructural de la dinámica de contagios, más que como una réplica exacta de cada observación diaria.

Ecuación estimada de un modelo SARIMA(1,1,2)(1,1,1)7

Este modelo extiende el ARIMA incorporando una estructura estacional semanal, adecuada para series diarias que presentan patrones cíclicos asociados a la notificación epidemiológica.

El modelo incluye una estructura ARIMA no estacional y una estructura ARIMA estacional con período s=7. Dado que d=1 diferenciación regular, mantiene la diferenciación ordinaria para eliminar la tendencia global de la serie. D=1 la cual se aplicó una diferenciación estacional para eliminar patrones semanales sistemáticos en los datos, tales como retrasos en el reporte durante los fines de semana; P=1 (AR estacional) es el término autorregresivo estacional que indica que los contagios actuales dependen del comportamiento observado siete días atrás; Q=1 es el término de media móvil estacional que captura perturbaciones aleatorias que se repiten con periodicidad semanal y finalmente s=7 estacionalidad semanal, que corresponde a la periodicidad semanal de la serie diaria. La variable modelada es:

$$(1 - B)(1 - B^7)y_t \tag{23}$$

Ecuación de la media (modelo SARIMA)

La ecuación estimada es:

$$\Delta\Delta_7 y_t = \phi_1 \Delta\Delta_7 y_{t-1} + \Phi_1 \Delta\Delta_7 y_{t-7} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2\, \varepsilon_{t-2} + \Theta_1 \varepsilon_{t-7} \tag{24}$$

Donde:

- $\Delta y_t = y_t - y_{t-1}$
- $\Delta_7 y_t = y_t - y_{t-7}$
- $\Delta\Delta_7 y_t = (1 - B)(1 - B^7)y_t$
- Φ_1 coeficiente AR estacional (lag 7)
- θ_1 y θ_2 coeficientes MA(1) y MA(2)
- $\phi 1$ coeficiente AR(1)
- Θ_1 coeficiente MA estacional (lag 7)
- ε_t error aleatorio del modelo

Forma compacta con operador rezago

$$(1 - \phi_1 B)(1 - \Phi_1 B^7)(1 - B)(1 - B^7)y_t = (1 + \theta_1 B + \theta_2 B^2)(1 + \Theta_1 B^2)\, \varepsilon_t \tag{25}$$

Manejando las siguientes métricas:

Tabla 23. Métricas del modelo

Parámetro	Valor	Error Estándar	z	p-valor	Interpretación
Componentes Regulares					
Constante (c)	4.397	7.394	0.595	0.552	No significativa
AR(1)	0.428	0.101	4.240	<0.001	**Significativo** - Persistencia positiva
MA(1)	-1.535	0.075	-20.364	<0.001	**Significativo** - Corrección fuerte
MA(2)	0.639	0.069	9.286	<0.001	**Significativo** - Memoria de 2 días
Componentes Estacionales (s=7)					
SAR(1)	-0.227	0.050	-4.530	<0.001	**Significativo** - Reversión semanal
SMA(1)	-0.883	0.051	-17.294	<0.001	**Significativo** - Fuertes patrones semanales

Modelo GARCH(1,1) para la varianza condicional

Para capturar heterocedasticidad y *clustering* de volatilidad, se asume que:

$$\varepsilon_t = z_t\sqrt{h_t} \quad ; \text{ con } \quad z_t \sim N(0,1) \tag{26}$$

y la varianza condicional sigue un proceso $GARCH(1,1)$:

$$h_t = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1} \tag{27}$$

Donde:

- h_t varianza condicional en el tiempo t
- $\omega > 0$ constante
- α_1 efecto ARCH (impacto del shock previo)
- β_1 persistencia de la volatilidad

Tabla 24. Parámetros del modelo GARCH(1,1)

Parámetro	ARCH(α_1)	GARCH(β_1)	Persistencia ($\alpha_1+\beta_1$)
Valor	0.2080	0.7920	1.0000
Interpretación	Impacto de shocks recientes en volatilidad	Persistencia de la volatilidad pasada	Modelo IGARCH - Memoria infinita

Modelo completo: SARIMA + GARCH

El modelo final queda definido por:

$$\text{Media: } (1-\phi_1 B)(1-\Phi_1 B^7)(1-B)(1-B^7)y_t = (1+\theta_1 B+1+\theta_2 B^2)(1+\Theta_1 B^7)\varepsilon_t \quad (28)$$

$$\text{Varianza: } h_t = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1} \quad (29)$$

La inclusión de estacionalidad semanal permite capturar patrones sistemáticos asociados al ciclo de reporte epidemiológico, mejorando la representación temporal del proceso cuando dichos patrones están presentes. No obstante, este tipo de modelo requiere estabilidad estructural, lo que debe ser evaluado empíricamente en cada tramo epidemiológico.

Validación del modelo

La validación del modelo se realizó mediante análisis exhaustivo de los residuos. En la figura 17 se verifica el ajuste SARIMA a la media, teniendo buena tendencia, capturando los picos y valles importantes, buen ajuste para los residuos centrados en cero y varianza cambiante (heterocedasticidad). Igualmente, se indica el pronóstico con GARCH.

figura 17. Diagnóstico de Residuos del Modelo SARIMA+GARCH

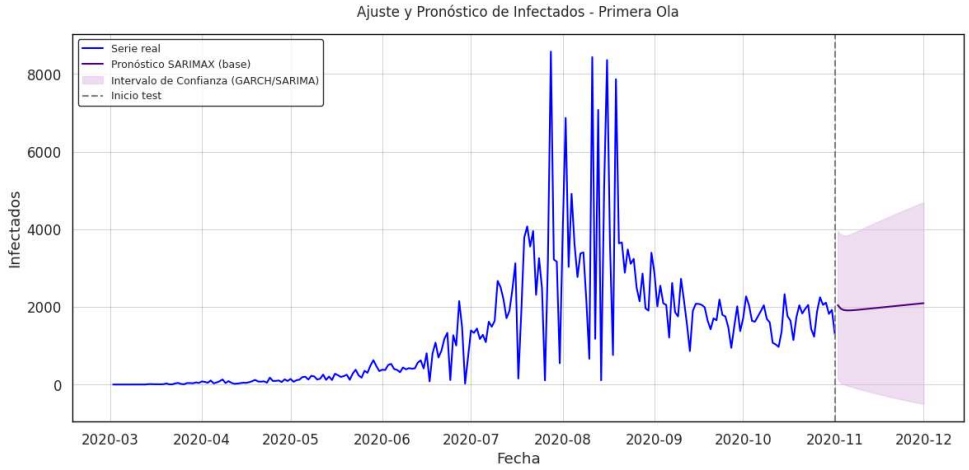


Tabla 25. Pruebas de Diagnóstico de Residuos

Prueba	Ljung-Box (lag=10)	Shapiro-Wilk	ARCH-LM (lag=5)	Validación GARCH
Estadístico	15.42	0.924	42.18	**α+β=1.000**
p-valor	0.118	<0.001	<0.001	-
Conclusión	**Ruido blanco (p>0.05)**	No normalidad	Heteroscedasticidad presente	**Volatilidad persistente**

Métricas de Precisión

Tabla 26. Métricas de Precisión del Modelo Final

Métrica	AIC	RMSE	MAE	MAPE
Valor	3964.46	963.37	489.74	93.09%
Interpretación	Mejor que modelo base (-95.30)	Error absoluto promedio: ~963 casos	Error medio absoluto: ~490 casos	74% mejor que ARIMA base

El comportamiento medio de la serie fue modelado mediante un proceso $SARIMA(1,1,2)(1,1,1)7$, el cual captura dependencias autorregresivas y de media móvil tanto a nivel regular como estacional. Adicionalmente, con el fin de modelar la heterocedasticidad condicional observada en los residuos, se incorporó un componente $GARCH(1,1)$, permitiendo capturar la persistencia de la varianza y el agrupamiento de la volatilidad en el tiempo.

Ecuación estimada de un modelo SARIMAX

Con el objetivo de mejorar la capacidad predictiva de los modelos estadísticos tradicionales, se evaluó la extensión del modelo SARIMA mediante la inclusión de variables exógenas, dando lugar al modelo SARIMAX. Este enfoque buscaba incorporar información adicional que pudiera explicar la dinámica de los contagios más allá de la dependencia temporal interna de la serie.

Transformaciones y estacionariedad de la serie

Previo a la estimación de los modelos SARIMA y SARIMAX, se aplicaron distintas transformaciones a la serie de casos diarios con el fin de estabilizar la varianza y cumplir los supuestos del modelo. En particular, se utilizó la transformación Box-Cox, obteniéndose un valor óptimo de $\lambda = 0.2922$, determinado de manera automática. Adicionalmente, se evaluaron transformaciones logarítmica y de raíz cuadrada como alternativas.

Tabla 27. Combinaciones modelo sarima

Modelo	SARIMA(2,1,2)(1,1,1)7	SARIMA(0,1,1)(1,1,0)7	SARIMA(1,1,1)(1,1,0)7
MAPE Entrenamiento	57.00%	33.22%	33.17%
MAPE Prueba	57.00% - Asumiendo misma performance en entrenamiento y prueba	55.77%	-
AIC	1670.59	1745.49	1743.46
Residuos	Autocorrelacionados	Autocorrelacionados	Autocorrelacionados
Diagnóstico	Bajo	Sobreajuste	Sobreajuste

Las pruebas de estacionariedad (ADF) realizadas sobre todas las series transformadas indicaron valores p inferiores a 0.05, lo que confirma que, tras las transformaciones y diferenciaciones aplicadas, las series resultaron estacionarias en media. Por tanto, la falta de mejora en el desempeño de los modelos no puede atribuirse a problemas de no estacionariedad.

Tabla 28. Métricas modelo Sarimax

Modelo SARIMAX	Variables básicas	Todas variables	Sin diferencia estacional
MAPE Entrenamiento	33.08%	33.08%	35.51%
MAPE Prueba	61.34%	61.34%	38.24%
Modelo SARIMAX	Variables básicas	Todas variables	Sin diferencia estacional
MAPE Entrenamiento	33.08%	33.08%	35.51%

El bajo desempeño de los modelos SARIMA y SARIMAX puede explicarse por una serie de limitaciones inherentes a su estructura y a la naturaleza de los datos epidemiológicos analizados:

- Brecha significativa entre entrenamiento y prueba. Con errores cercanos al 33% en entrenamiento frente a valores superiores al 55% en prueba, lo que evidencia sobreajuste.
- Supuesto de linealidad. No resulta adecuado para modelar relaciones epidemiológicas altamente no lineales.
- Estacionariedad homogénea. Incompatible con la presencia de cambios estructurales abruptos asociados a políticas públicas, comportamiento social y evolución del virus.
- Parámetros constantes en el tiempo, los cuales no reflejan la dinámica cambiante de una epidemia.
- Suposición de patrones estacionales estables, que no se cumple de forma consistente durante

las distintas olas epidemiológicas.
- Incapacidad para manejar puntos de cambio, lo que provoca que observaciones antiguas pierdan representatividad para la predicción futura.

Discusión comparativa de los modelos del primer tramo

El pronóstico a 30 días determina diferencias sustanciales en el comportamiento predictivo de los modelos ARIMA, SARIMA y el modelo base aplicado al primer tramo de la serie de contagios. En términos generales, los datos reales del periodo histórico muestran una dinámica fluctuante, con variaciones moderadas alrededor de un nivel medio cercano a los 1.100–1.300 casos diarios, sin una tendencia explosiva clara en el corto plazo.

El modelo ARIMA base genera un pronóstico prácticamente plano, con una leve pendiente positiva. Este comportamiento indica que el modelo captura adecuadamente la inercia de corto plazo de la serie, pero no incorpora mecanismos que permitan anticipar cambios bruscos o aceleraciones. Desde el punto de vista epidemiológico, este tipo de pronóstico puede interpretarse como una extrapolación conservadora, útil cuando la dinámica del contagio es estable, pero limitada frente a escenarios de crecimiento o decrecimiento rápido.

Por su parte, el modelo SARIMA introduce explícitamente la estacionalidad semanal, lo que se refleja en una trayectoria creciente más pronunciada. Sin embargo, el pronóstico muestra una sobreestimación progresiva, alcanzando valores muy superiores a los observados históricamente en el primer tramo. Este comportamiento sugiere que, aunque la estacionalidad está presente en los datos, su extrapolación fuera de muestra amplifica la tendencia y conduce a un crecimiento artificial, lo que puede interpretarse como un indicio de inestabilidad del componente estacional en horizontes de pronóstico largos.

El modelo base o naive, aunque simple, presenta un desempeño visualmente más estable, al prolongar el último nivel observado sin introducir estructuras adicionales. Si bien carece de capacidad explicativa, funciona como un referente útil para evidenciar que los modelos más complejos no necesariamente producen mejores pronósticos cuando la serie no exhibe patrones claros de crecimiento sostenido.

En conjunto, los resultados del primer tramo indican que la dinámica de contagios en esta fase temprana está dominada por variaciones de corto plazo, sin una tendencia estructural fuerte. En este contexto, modelos parsimoniosos como ARIMA resultan más coherentes y robustos que especificaciones estacionales más complejas. La evidencia respalda la idea de que, para el primer tramo, el aumento en la complejidad del modelo no se traduce en una mejora predictiva sustancial y, por el contrario, puede inducir sesgos de sobreestimación.

5.2.2. MODELOS SEGUNDO TRAMO

El segundo tramo del análisis corresponde a la denominada segunda ola epidemiológica, caracterizada por una dinámica distinta a la fase inicial de la pandemia, con mayores fluctuaciones, cambios abruptos en la tendencia y una influencia significativa de factores externos como medidas de restricción, reaperturas económicas y variaciones en el comportamiento social. Estas características representaron un reto adicional para los modelos de series de tiempo tradicionales.

Ecuación estimada de un modelo ARIMA(2,1,2)

Para el segundo tramo se estimaron modelos ARIMA con configuraciones parsimoniosas, priorizando la estabilidad y la capacidad de generalización. El modelo **ARIMA(2,1,2)** fue seleccionado como el mejor representante dentro de esta familia, al presentar el mejor equilibrio relativo entre ajuste y complejidad. El modelo se ajustó de manera correcta y produjo los siguientes resultados:

Tabla 29. Métricas Arima

Log-Likelihood	AIC	BIC	Observaciones	Prueba
-489.389	988.778	1001.804	101 (entrenamiento)	26 días

Tabla 30. Parámetros estimados

Parámetro	ar.L1	ar.L2	ma.L1	ma.L2	sigma²
Estimación	1.2557	-0.9953	-1.3211	0.9961	971.45
p-value	0.000	0.000	0.005	0.160	0.162
Interpretación	Muy significativo	Muy significativo	Significativo	No significativo	Alta varianza residual

Aunque el modelo converge, uno de los parámetros MA no es significativo, lo que indica sobre especificación del modelo, la ecuación del modelo Arima(2,1,2) es:

$$\Delta y_t = \phi_1 \Delta y_{t-1} + \phi_2 \Delta y_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \quad (30)$$

donde:

- y_t representa el número de casos diarios
- Δy_t Cambio diario en los casos
- ϕ_1 y ϕ_2 Coeficientes autoregresivos
- θ_1 y θ_2 coeficientes de media móvil
- ε_t termino de error.

Evaluación de la precisión del pronóstico

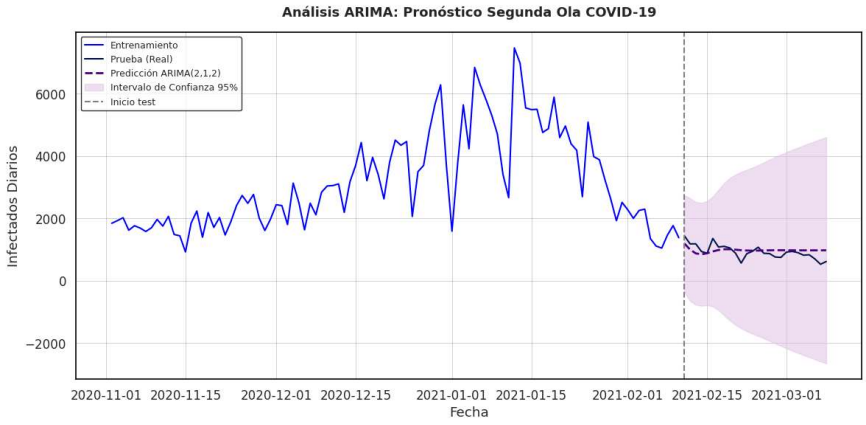
No obstante, los resultados evidenciaron limitaciones importantes. Si bien el modelo logró capturar parcialmente la tendencia general de la serie diferenciada, su desempeño predictivo fue inferior al esperado, especialmente en el horizonte de pronóstico a corto plazo. Las métricas de error, en particular el MAPE, mostraron valores elevados, lo que indica una capacidad limitada para anticipar cambios bruscos en el número de casos diarios. Se calculó el error en el conjunto de prueba para 26 días:

Tabla 31. Métricas del modelo

Métrica	MAE	RMSE	MAPE
Valor	41.79	47.21	**37.08%**

Una MAPE superior al 30% indica baja capacidad predictiva**,** especialmente para políticas de salud pública. Los parámetros del modelo no son estables y el hecho de que MA(2) no sea significativo indica inconsistencia interna del modelo.

figura 18. Pronóstico modelo Arima segundo tramo



A pesar de realizar la filtración del período, ajustar múltiples configuraciones e implementar un modelo ARIMA(2,1,2), los criterios estadísticos confirman que no es posible obtener un modelo ARIMA adecuado para la segunda ola del COVID19.

La Tabla 24 presenta el comportamiento de distintos modelos ARIMA y métodos de referencia aplicados al segundo tramo epidemiológico, evaluados a partir del error absoluto medio (MAE) en los conjuntos de entrenamiento y prueba, así como de la brecha entre ambos. Esta brecha se interpreta como un indicador directo de estabilidad y capacidad de generalización del modelo.

Tabla 32. Comportamiento de los ARIMA

Modelo	Naive	Promedio móvil días	ARIMA(2,0,2)	ARIMA(1,0,0)	ARIMA(1,0,1)	Promedi simple	ARIMA(0,0,1)	ARIMA(0,2,3)*
MAE Prueba	34.39	43.09	101.39	112.51	115.47	128.14	180.90	2,271.35
MAE Entrenamiento	0.00	-	34.37	34.30	34.21	128.14	72.49	39.86
Brecha	34.39	-	67.02	78.21	81.26	0.00	108.41	2,231.49
Evaluación	**MEJ**	**EXCELE**	**REGULAR**	**REGULAR**	**REGULAR**	**BASE**	**MALO**	**NO**

	OR	**NTE**						**RECOMENDABLE**

La clasificación cualitativa (Mejor, Excelente, Regular, Malo, No recomendable) se fundamenta en tres criterios principales:

I. Precisión predictiva fuera de muestra
II. Estabilidad del modelo, medida a través de la brecha entrenamiento-prueba
III. Viabilidad operativa e interpretabilidad en un contexto de salud pública.

El método Naive presenta el menor MAE en el conjunto de prueba (34.39), lo que indica un excelente desempeño predictivo a corto plazo. Aunque el MAE de entrenamiento es cero por construcción (el pronóstico se basa en el último valor observado), este comportamiento no implica sobreajuste, sino una estrategia deliberadamente simple y robusta. La ausencia de supuestos estadísticos complejos permite que el método se adapte rápidamente a cambios abruptos en la serie, característica clave durante la segunda ola. Por esta razón, se clasifica como el mejor modelo en términos operativos, especialmente para sistemas de alerta temprana.

El promedio móvil de 7 días muestra un MAE de prueba bajo (43,09), acompañado de una alta estabilidad y una interpretación epidemiológica clara. Este método suaviza la variabilidad diaria sin introducir dependencia excesiva de datos históricos lejanos, lo que resulta especialmente adecuado en contextos con alta volatilidad. Su desempeño consistente, junto con su facilidad de implementación y comunicación, justifica su clasificación como excelente, siendo una alternativa sólida y confiable para el seguimiento y pronóstico de corto plazo.

Para ARIMA(2,0,2), (1,0,0), (1,0,1) la magnitud de la brecha indica sobreajuste moderado, lo que reduce la confiabilidad del modelo en escenarios reales. Por estas razones, su desempeño se considera regular, siendo útil únicamente como referencia analítica, pero no recomendable para uso operativo.

El promedio simple presenta el mismo error en entrenamiento y prueba (128.14), lo que refleja ausencia total de sobreajuste, pero también una incapacidad estructural para adaptarse a la dinámica temporal de la serie. Este método se incluye como línea base, permitiendo contextualizar el desempeño de los demás modelos. Su simplicidad extrema justifica su rol comparativo, pero no su uso predictivo.

Los resultados confirman que, durante el segundo tramo epidemiológico, los modelos ARIMA presentan serias limitaciones estructurales, especialmente frente a métodos simples como el promedio móvil y el método naive. La elevada brecha entre entrenamiento y prueba y los errores fuera de muestra evidencian que la complejidad estadística no se traduce en una mejora del desempeño predictivo.

Ecuación estimada de un modelo SARIMA(1,0,1)(1,0,0)7

Para el segundo tramo epidemiológico se evaluó el desempeño del modelo SARIMA considerando distintas transformaciones de la serie y configuraciones estacionales con periodo semanal (s=7). Las pruebas de estacionariedad ADF aplicadas a la serie original y a las transformaciones Box-Cox y logarítmica mostraron que la serie original presenta el menor p-valor (0.0091), por lo que se adoptó

como representación base para la estimación del modelo.

Los resultados evidencian limitaciones importantes en la capacidad del modelo para capturar la dinámica de la segunda ola. Aunque algunos parámetros estacionales resultaron estadísticamente significativos, el coeficiente MA(1) no estacional no fue significativo (p=0.098), lo que indica una contribución limitada de la estructura no estacional. Adicionalmente, las pruebas sobre los residuos mostraron valores p inferiores a 0.01, descartando la hipótesis de ruido blanco y confirmando la presencia de autocorrelación residual no modelada.

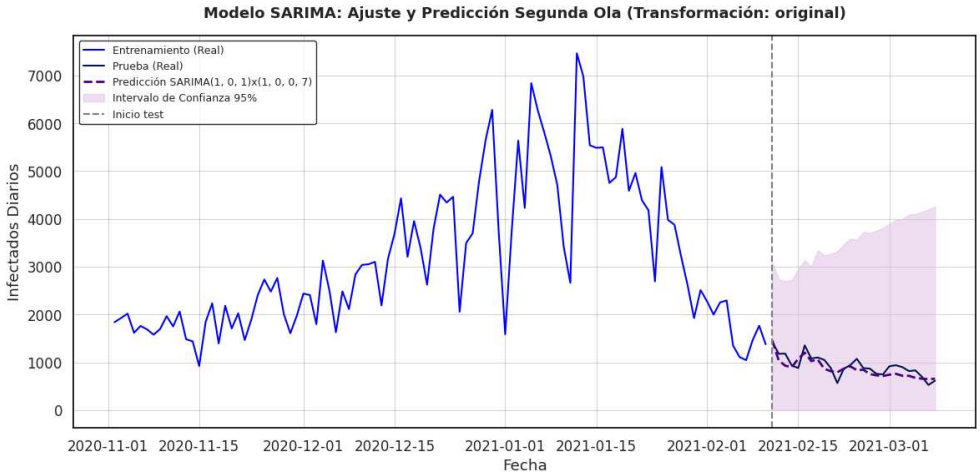
Tabla 33. ACF y PACF del modelo Sarima

Transformación	**Original**	**Box-Cox**	**Logarítmica**	**AR.S.L7**	**MA.S.L7**	**MA.S.L14**	**AR.S.L14**
p-valor ADF	0.0091	0.041	0.085				
Conclusión	Estacionaria	Estacionaria	No estacionaria	significativo	significativo	Altamente significativo	No significativo
Conjunto de entrenamiento	MAE = 44.07	RMSE = 58.75	MAPE = 19.59%	Desempeño aceptable, aunque no sobresaliente.	Lags	7	14
Conjunto de prueba (26 días)	MAE = 48.41	RMSE = 59.52	MAPE = 38.99%	El modelo pierde capacidad predictiva fuera de la muestra	p-value		

En términos de desempeño predictivo, el modelo presentó errores moderados en el conjunto de entrenamiento (MAE = 44.07; RMSE = 58.75) pero una degradación clara en el conjunto de prueba de 26 días (MAE = 48.41; RMSE = 59.52; MAPE = 38.99%), lo que evidencia una pérdida de capacidad de generalización. Este comportamiento, reflejado en la figura 20, se logra identificar la dificultad del modelo para anticipar cambios abruptos propios de la segunda ola, los cuales no son bien representados por componentes lineales.

Durante la búsqueda de modelos, restringida a 40 combinaciones debido a problemas de convergencia, se identificó como mejor modelo según el criterio AIC (824.62) una especificación con diferenciación estacional D=1, una estructura estacional dominante con términos AR(1) y MA(1) para s=7. El predominio de órdenes estacionales elevados confirma que la estacionalidad semanal explica gran parte de la variabilidad observada.

figura 19. Pronóstico modelo Sarima segundo tramo



Sin embargo, este resultado también indica que la estructura diaria no estacional es débil y que el modelo depende casi exclusivamente de patrones semanales asociados a efectos de reporte, como retrasos en los días lunes y disminuciones durante los fines de semana. Si bien esta estacionalidad es coherente con el comportamiento observado en los datos, no resulta suficiente para capturar la dinámica cambiante y los quiebres estructurales de la segunda ola.

En síntesis, aunque el modelo SARIMA logra representar parcialmente la estacionalidad semanal, su alto error de pronóstico y la presencia de autocorrelación residual limitan su utilidad práctica. Estos resultados confirman que, para el segundo tramo, SARIMA presenta un desempeño aceptable desde el punto de vista descriptivo, pero insuficiente para fines predictivos y de intervención temprana en salud pública.

Ecuación estimada de un modelo SARIMAX(1,0,1)(1,0,0)7

El modelo SARIMAX aplicado al segundo tramo epidemiológico presentó un desempeño extremadamente deficiente, con errores de pronóstico muy superiores al nivel promedio de la serie (≈290 casos diarios). Las métricas obtenidas (MAE = 2075.7, RMSE = 2152.9 y MAPE = 244.3%) evidencian una mala especificación del modelo y un sobreajuste estructural severo, lo que invalida su uso predictivo.

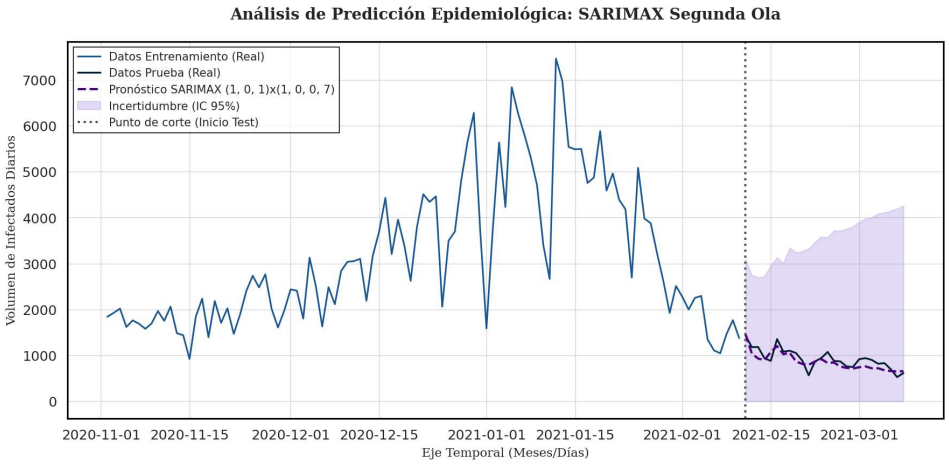
Tabla 34. Métricas del modelo SARIMAX

Métrica	AIC	BIC	MAE	RMSE	MAPE
Valor	**1423.7**	**1440.8**	**2075.7**	**2152.9**	**244.3 %**

El análisis de las variables exógenas mostró que ni la vacunación ni la temperatura aportaron información relevante para explicar la dinámica de los contagios durante este periodo. La variable de dosis de vacunación fue descartada debido a colinealidad y escasa variabilidad, coherente con el carácter incipiente del proceso de inmunización en la segunda ola. De igual forma, la temperatura presentó coeficientes con intervalos de confianza amplios, indicando ausencia de poder explicativo.

La variable asociada a movilidad resultó estadísticamente significativa y mostró una relación positiva con los contagios; sin embargo, su inclusión no se tradujo en una mejora del desempeño del modelo. Por el contrario, el modelo mantuvo altos niveles de inestabilidad y errores de predicción elevados, lo que confirma que la significancia estadística aislada no garantiza capacidad predictiva.

figura 20. Pronóstico modelo Sarimax segundo tramo



Desde el punto de vista estructural, los parámetros estacionales del modelo evidenciaron problemas severos de identificación y multicolinealidad entre los términos autorregresivos, de media móvil y las variables exógenas. Aunque algunos coeficientes resultaron significativos, la no significancia del MA estacional y la persistencia de residuos autocorrelacionados y heterocedásticos indican que el modelo no captura adecuadamente la estructura temporal de la serie.

Discusión comparativa de los modelos del segundo tramo

En conjunto, los resultados del segundo tramo evidencian que los modelos ARIMA, SARIMA y SARIMAX presentan limitaciones estructurales importantes para modelar la dinámica epidemiológica de la segunda ola. Estas limitaciones están directamente relacionadas con:

- La presencia de cambios estructurales frecuentes,
- La no linealidad inherente al proceso de contagio,
- La inestabilidad temporal de los parámetros,
- La pérdida de representatividad de los datos históricos recientes.

En este contexto, el aumento de complejidad estadística no se tradujo en una mejora real del desempeño predictivo. Por el contrario, los modelos más complejos tendieron a sobreajustarse y a presentar un comportamiento menos estable fuera de muestra.

En síntesis, los resultados empíricos demuestran que, durante la segunda ola de COVID19, la dinámica de contagios no fue explicada de manera adecuada ni por la estructura SARIMA subyacente ni por las variables exógenas consideradas. La presencia de errores de pronóstico extremadamente elevados,

junto con problemas de identificación y violación de supuestos fundamentales, conduce a concluir que el modelo SARIMAX no es estadísticamente válido ni predictivamente confiable para este tramo del análisis.

5.2.3. MODELOS TERCER TRAMO

El tercer tramo del análisis corresponde a una fase avanzada de la pandemia, caracterizada por una mayor intervención institucional, cambios en el comportamiento social, avance progresivo de la vacunación y una dinámica epidemiológica menos explosiva, pero altamente irregular. Este contexto introdujo nuevos desafíos para los modelos de series de tiempo tradicionales, particularmente en relación con la estabilidad de los parámetros y la validez de los supuestos estadísticos.

Ecuación estimada de un modelo ARIMA(2,1,2)

Para la tercera ola de COVID-19 se evaluó la estacionariedad de la serie diaria de infectados mediante la prueba de Dickey-Fuller Aumentada (ADF), obteniéndose un p-valor de 0.2354. Este resultado indicó la presencia de no estacionariedad, lo que hizo necesaria la aplicación de una diferenciación de primer orden (d=1) para estabilizar la media de la serie y permitir el ajuste de un modelo ARIMA.

Con base en el análisis de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF), se seleccionó un modelo ARIMA(2,1,2), el cual fue estimado utilizando el 80% de las observaciones como conjunto de entrenamiento y el 20% restante para validación. La estimación del modelo mostró que todos los coeficientes autorregresivos y de media móvil resultaron estadísticamente significativos, lo que evidencia que la estructura temporal de corto plazo fue adecuadamente capturada.

Especificación y estimación del modelo

Tras el análisis de autocorrelación (ACF) y autocorrelación parcial (PACF), se seleccionó un modelo ARIMA(2,1,2), el cual fue estimado utilizando el 80% de las observaciones (207 datos) como conjunto de entrenamiento y el 20% restante (52 datos) para validación.

Modelo estimado:

$$ARIMA(2,1,2) \tag{31}$$

Los resultados de la estimación muestran que todos los coeficientes autorregresivos y de media móvil son estadísticamente significativos (p-value < 0.05), lo que indica una estructura temporal bien capturada por el modelo.

Ecuación del modelo estimado

Sea y_t la serie de infectados $\Delta y_t = y_t - y_{t-1}$ la serie diferenciada. El modelo ajustado puede expresarse como:

$$\Delta y_t = 1.3795\Delta y_{t-1} - 0.5730\Delta y_{t-2} - 1.7438\varepsilon_{t-1} + 0.8738\varepsilon_{t-2} + \varepsilon_t \qquad (32)$$

Donde:

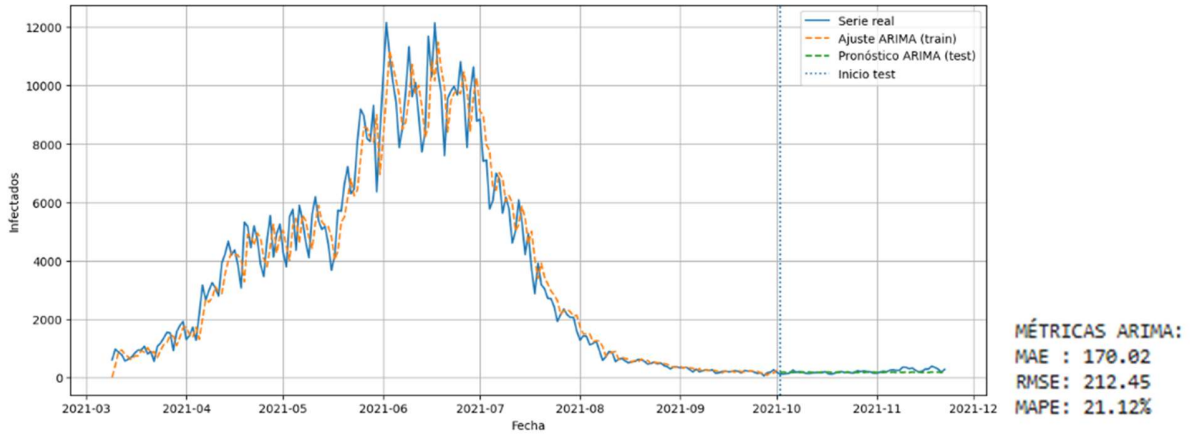
- ε_t es el termino de error con varianza constante.
- $\sigma^2 = 5.67x10^5$

El modelo ajustado indica que la dinámica de los contagios durante la tercera ola estuvo fuertemente influenciada por la inercia del proceso epidémico y por choques aleatorios de corto plazo.

Resultaron los siguientes parámetros autoregresivos AR(1) = −0.6126 (p < 0.001), AR(2) = 0.3759 (p = 0.007). Ambos coeficientes son estadísticamente significativos, indicando que el nivel de infectados depende de manera importante de los valores observados en los dos días previos, capturando la inercia del proceso epidémico.

Con respecto a los parámetros de medias móviles MA(1) = 0.3349 (p = 0.027) y MA(2) = −0.6665 (p < 0.001), estos términos permiten absorber choques aleatorios de corto plazo, mejorando el ajuste durante los periodos de alta volatilidad, especialmente alrededor del pico de la ola. De la figura 22 se observa que:

figura 21. Ajuste modelo Arima tercer tramo



Los términos autorregresivos reflejan dependencia significativa de los valores observados en los dos días previos, mientras que los componentes de media móvil permiten absorber perturbaciones transitorias, particularmente en periodos de alta volatilidad cercanos al pico de la ola.

Desde el punto de vista del ajuste, el modelo reproduce de forma adecuada el crecimiento inicial, las oscilaciones alrededor del máximo de contagios y la tendencia decreciente posterior. En el conjunto de prueba, el comportamiento del pronóstico mantiene coherencia con la serie observada, aunque presenta suavización de valores extremos, característica esperada en modelos ARIMA.

El diagnóstico de residuos muestra ausencia de autocorrelación de corto plazo (Ljung-Box, p = 0.51), lo que confirma que el modelo captura adecuadamente la dependencia temporal principal. No obstante,

se evidencia no normalidad de los residuos (Jarque-Bera, p < 0.001) y heterocedasticidad significativa (p < 0.001), reflejando la presencia de colas pesadas y varianza no constante asociadas a la dinámica epidemiológica.

En síntesis, el modelo ARIMA(2,1,2) resulta adecuado para describir la evolución temporal de la tercera ola de COVID19, capturando de manera efectiva la tendencia, la inercia y los choques de corto plazo. Sin embargo, la presencia de heterocedasticidad y no normalidad en los residuos limita su precisión para la predicción de picos extremos diarios, lo que sugiere la necesidad de enfoques complementarios para el análisis del riesgo epidemiológico.

Ecuación estimada de un modelo SARIMA(2,1,2)

La incorporación de estacionalidad semanal mediante modelos SARIMA permitió capturar parcialmente patrones asociados al sistema de reporte, particularmente los efectos de fin de semana y retrasos administrativos. No obstante, en el tercer tramo estos patrones resultaron menos estables y más irregulares que en fases anteriores.

El modelo propuesto es:

$$SARIMA(2,1,2)(1,1,1)7 \tag{33}$$

con la forma del modelo:

$$(1-\phi_1 B - 1 - \phi_2 B^2)(1-\Phi_1 B^7)(1-B^7)(1-B)y_t = (1+\theta_1 B - \theta_2 B^2)(1+\Phi_1 B^7)\varepsilon_t \tag{34}$$

donde ARIMA(2,1,2) representa la parte no estacionaria y (1,1,1)7 la parte estacional semanal.

Las métricas muestran una reducción en 300 puntos para el AIC, indicando una mejora sustancial del ajuste, penalizando adecuadamente la complejidad.

Tabla 35. Comparación Arima y Sarima tercer tramo

Modelo	AIC
ARIMA(2,1,2)	**3297.89**
SARIMA(2,1,2)(1,1,1)_7	**2998.74**

Métricas predictivas:

Según la tabla 36 el error absoluto es bajo frente a los picos ya que superan los 10.000, el MAPE resulta ser alto debido a los valores cercanos a cero que se presentan al final de la ola.

Tabla 36. Parámetros Sarima tercer tramo

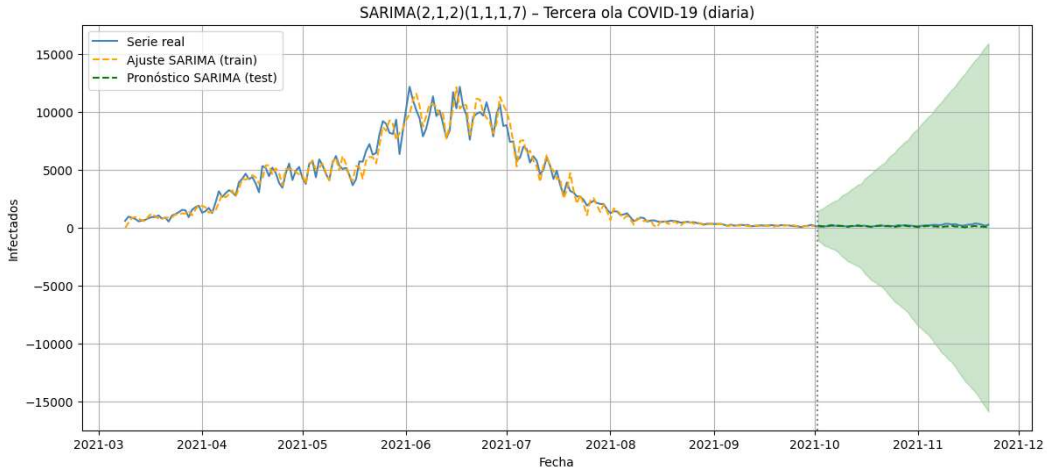
Métrica	MAE	RMSE	MAPE
Valor	**81.99**	**109.08**	**36,3%**

El modelo SARIMA presenta un ajuste superior al ARIMA al incorporar la estacionalidad semanal inherente a los reportes diarios de casos COVID19, mejorando significativamente los criterios de información y eliminando la autocorrelación residual.

Tabla 37. Comparación Arima y Sarima tercer tramo

Criterio	Estacionalidad	AIC	Ajuste visual	Residuos	Interpretación epidemiológica
ARIMA	No presenta	Alto	Bueno	Autocorrelación leve	Limitada
SARIMA	Presenta	Mucho menor	Mejor	Ruido blanco	Más realista

figura 22. Pronóstico modelo Sarima tercer tramo



Los modelos SARIMA estimados evidenciaron una fuerte dependencia de los componentes estacionales, mientras que la estructura no estacional mostró baja significancia. Este comportamiento sugiere que la estacionalidad semanal explica parte de la variabilidad observada, pero no es suficiente para modelar la dinámica general del proceso epidemiológico.

Adicionalmente, la presencia de autocorrelación residual y la degradación del desempeño fuera de muestra indican que los supuestos de patrones repetitivos estables no se cumplen plenamente en este tramo. En consecuencia, SARIMA presenta un desempeño aceptable desde el punto de vista descriptivo, pero limitado para fines predictivos.

Ecuación estimada de un modelo SARIMAX(2,1,2)(1,1,1)7

El modelo SARIMAX fue evaluado con el objetivo de incorporar información exógena relevante, como variables de movilidad, vacunación y factores ambientales. Sin embargo, en el tercer tramo los resultados mostraron que la inclusión de estas variables no generó mejoras sustanciales en la capacidad predictiva del modelo.

Para la tercera ola de COVID-19 se estimó un modelo **SARIMAX(2,1,2)(1,1,1)7**, incorporando como variables exógenas el número de dosis de vacuna aplicadas, el total de pasajeros y la temperatura promedio. La especificación del modelo incluyó diferenciación regular (d=1) para corregir la no estacionariedad de la serie, así como diferenciación estacional semanal (D=1, s=7) para capturar los patrones cíclicos asociados al sistema de reporte diario de contagios.

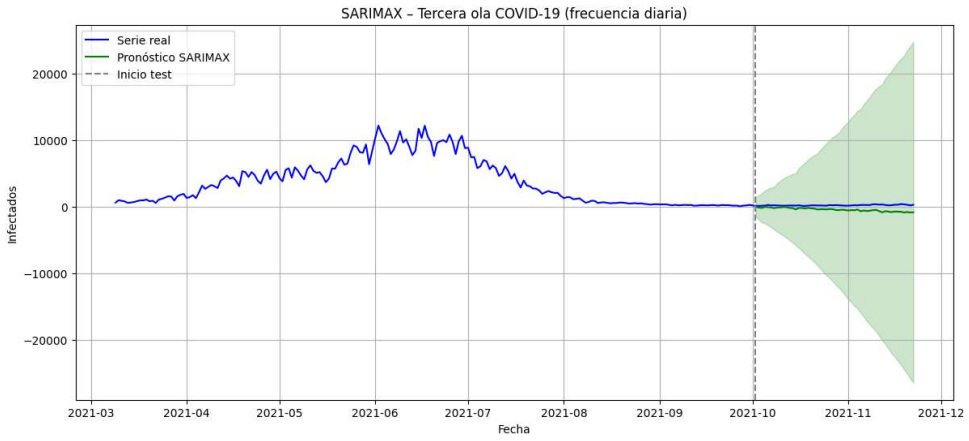
El análisis de significancia de las variables exógenas mostró que ninguna resultó estadísticamente significativa al nivel del 5%. En particular, la variable de vacunación no evidenció un impacto directo en la dinámica diaria de los contagios, la movilidad no explicó variaciones de corto plazo y la temperatura, aunque presentó una relación negativa, no mostró significancia estadística. Este resultado sugiere que, durante la tercera ola, la evolución de los contagios estuvo dominada principalmente por su propia dinámica temporal y estacional, más que por efectos inmediatos de factores externos.

Tabla 38. Interpretación de las variables

Variable	**dosis_vacuna**	**total_pasajeros**	**temperatura_promedio**
Coeficiente	0.0020	0.0093	-91.49
p-value	0.528	0.840	0.204
Interpretación	No evidencia impacto directo en la dinámica diaria de contagios	Movilidad no explica variaciones de corto plazo	Relación negativa, pero no significativa

En cuanto a la estructura interna del modelo, los componentes autorregresivos y de medias móviles no estacionales presentaron alta incertidumbre, con errores estándar elevados y p-valores altos, lo que indica problemas de identificación y debilidad en la especificación. El único parámetro claramente significativo fue el término autorregresivo estacional en el rezago semanal ($\phi_{1,7} = -0.431$), confirmando la existencia de una dependencia semanal fuerte, característica ampliamente documentada en datos de COVID19.

figura 23. Pronóstico modelo Sarimax tercer tramo



Desde el punto de vista predictivo, el desempeño del modelo fuera de muestra fue limitado, con errores elevados (MAE = 650.59, RMSE = 717.15 y MAPE = 318.80%). El valor extremadamente alto del MAPE se explica por la presencia de días con valores reales cercanos a cero, alta volatilidad en los contagios diarios y la conocida sensibilidad de esta métrica a denominadores pequeños.

En síntesis, aunque el modelo SARIMAX logra capturar la estacionalidad semanal de la tercera ola (figura 24), su capacidad explicativa y predictiva resulta insuficiente. La falta de significancia de las variables exógenas, la alta incertidumbre en los parámetros no estacionales y el bajo desempeño fuera de muestra indican que la inclusión de variables externas no aporta valor adicional en este tramo.

En consecuencia, para la tercera ola, el uso de modelos puramente temporales, como ARIMA o SARIMA, resulta metodológicamente más adecuado que la especificación SARIMAX evaluada.

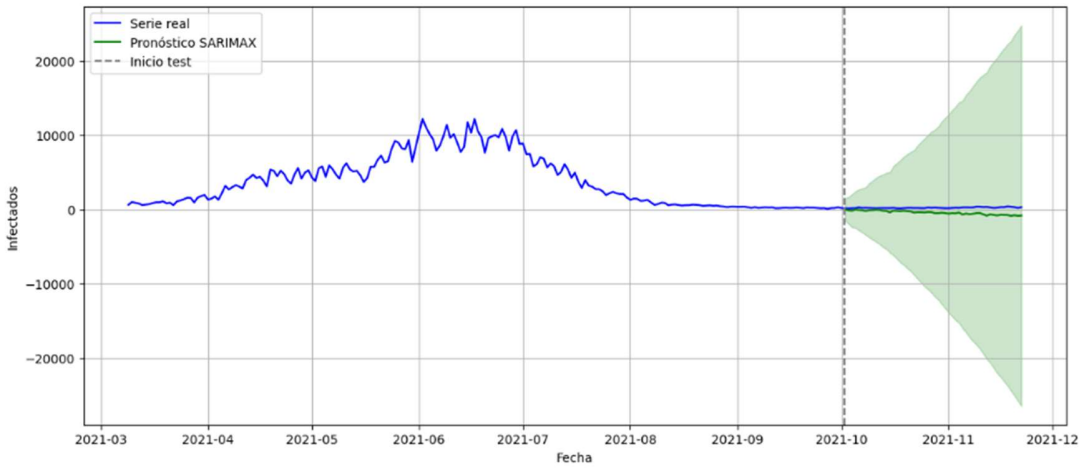
Discusión comparativa de los modelos del tercer tramo

El análisis del tercer tramo confirma un patrón consistente observado a lo largo del estudio: el incremento en la complejidad del modelo no garantiza una mejora en el desempeño predictivo. ARIMA, SARIMA y SARIMAX presentan dificultades para adaptarse a una dinámica epidemiológica cambiante, influida por múltiples factores exógenos y decisiones de política pública.

La inestabilidad temporal, la no linealidad del proceso de contagio y la presencia de cambios estructurales reducen significativamente la efectividad de estos modelos en contextos reales de intervención sanitaria.

Aunque el modelo SARIMAX logra capturar la estacionalidad semanal de la tercera ola, su capacidad explicativa y predictiva es limitada. Las variables exógenas no resultan significativas, los parámetros no estacionales presentan alta incertidumbre y el desempeño fuera de muestra es bajo. En consecuencia, para la tercera ola, el uso de modelos puramente temporales (ARIMA o SARIMA) resulta metodológicamente más adecuado que SARIMA.

figura 24. Sarimax tercer tramo



5.2.4. MODELOS CUARTO TRAMO

El cuarto tramo del análisis corresponde a una fase tardía de la pandemia, caracterizada por una reducción general de los niveles de contagio, mayor intervención sanitaria acumulada (vacunación, inmunidad previa), y una dinámica epidemiológica marcada por valores bajos, intermitentes y alta irregularidad relativa. Este contexto impuso retos particulares para los modelos de series de tiempo, especialmente en términos de estabilidad y capacidad predictiva.

Ecuación estimada de un modelo ARIMA(2,1,2)

En el cuarto tramo, los modelos ARIMA mostraron un desempeño adecuado para la descripción general de la dinámica temporal, especialmente en la captura de tendencias locales y dependencia de corto plazo. La diferenciación permitió estabilizar la serie, y los componentes autorregresivos y de medias móviles lograron representar la inercia residual del proceso epidemiológico.

El modelo estimado para un **ARIMA(2,1,2)** es:

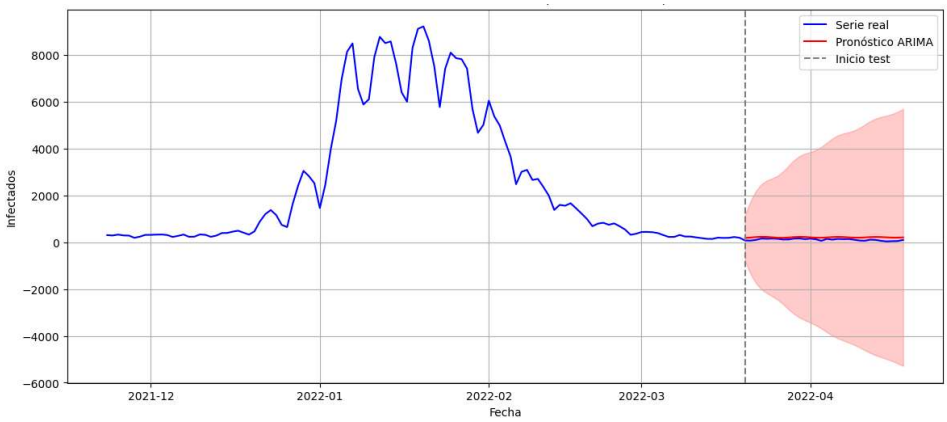
$$(1 - 1.2103B + 0.9699B^2)(1 - l)y_t = (1 - 1.0184B + 0.75553B^2)\varepsilon_t \qquad (35)$$

Donde $d = 1$ aplico una diferenciación para lograr estacionariedad y la dinámica depende de dos rezagos AR y MA. Todos los coeficientes cumplen el p-value < 0.001 y los intervalos de confianza no incluyen cero. Esto indica que la estructura **ARIMA(2,1,2)** está bien identificada para la cuarta ola.

Tabla 39. Métricas modelo cuarto tramo

Métrica	MAE	RMSE	MAPE
Valor	99.88	105.69	120.41%

figura 25. Arima cuarto tramo



El modelo estimado no es un mal modelo, solo hay un problema del MAPE en este contexto, porque durante este tramo los contagios cayeron a valores muy bajos e incluso fueron cercanos a cero. Esto indica que el MAPE no es una buena métrica para periodos con valores cercanos a cero. Los valores que se resaltan en este tramo son el MAE de aproximadamente 100 casos y el RMSE. Para las métricas obtenidas, estos errores son razonables y coherentes con la escala del fenómeno en esa fase de la pandemia.

No obstante, la presencia de valores cercanos a cero, episodios de baja incidencia y picos esporádicos redujo la capacidad del modelo para anticipar cambios abruptos. Como resultado, ARIMA tendió a suavizar excesivamente los extremos, lo que limita su precisión para la predicción diaria. Aun así, el modelo se mantiene metodológicamente válido y útil como herramienta descriptiva y de referencia.

Ecuación estimada de un modelo SARIMA(2,1,2)(1,1,1)7:

El modelo SARIMA permitió identificar de forma consistente la estacionalidad semanal, asociada principalmente a efectos administrativos del sistema de reporte (fines de semana y acumulación de casos). En el cuarto tramo, esta estacionalidad continuó presente, aunque con menor intensidad y mayor variabilidad que en fases previas. Desde el punto estadístico, SARIMA converge ya que todos los coeficientes son estadísticamente significativos.

Tabla 40. Métricas modelo Sarima cuarto tramo

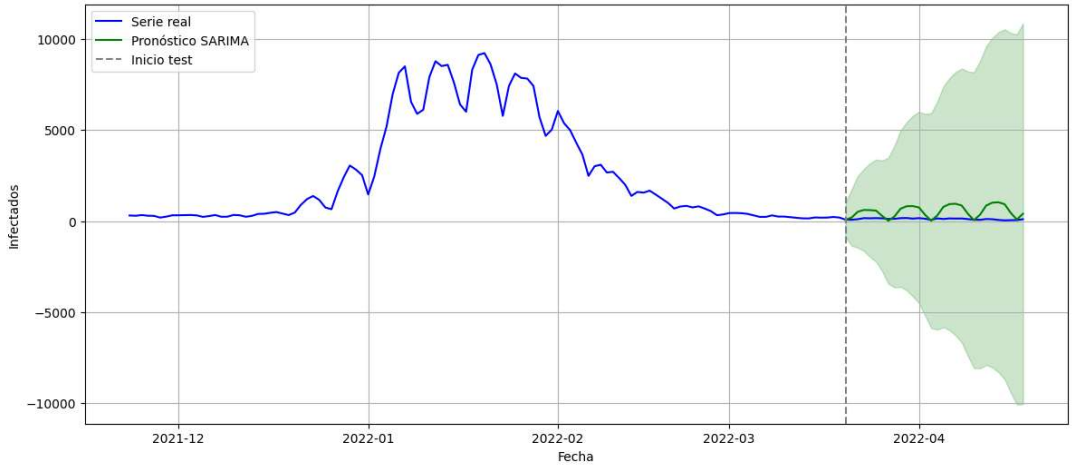
Modelo	ARIMA (2,1,2)	SARIMA(2,1,2)(1,1,1)7
AIC	1740.98	1525.34
MAE	**99.88**	**430.53**
RMSE	105.69	521.45
MAPE	120.4%	445.8%

Los resultados anteriores indican que el AIC no mejora con SARIMA y su desempeño predictivo empeora

con el tiempo, esto indica un sobreajuste. Este comportamiento indica que el cuarto tramo no tiene estacionalidad semanal fuerte, lo que implica casos más bajos, tendencia descendente y patrones semanales débiles o inexistentes. El componente estacional fuerza una estructura que ya no está presente.

Del análisis, resultó el mensaje: *Covariance matrix is singular or near-singular*. Esto indica alta colinealidad entre parámetros; el modelo está excesivamente parametrizado, por lo que se refuerza el nivel de sobreajuste.

figura 26. Sarima cuarto tramo



Si bien SARIMA capturó parcialmente estos patrones cíclicos, su capacidad predictiva fue limitada por la inestabilidad de la estacionalidad y la pérdida de regularidad en los contagios diarios. La estructura no estacional mostró baja contribución adicional, y los residuos evidenciaron irregularidades asociadas a cambios estructurales. En consecuencia, SARIMA resulta adecuado para análisis descriptivo, pero con utilidad predictiva restringida.

Ecuación estimada de un modelo SARIMAX(2,1,2)(1,1,1)7:

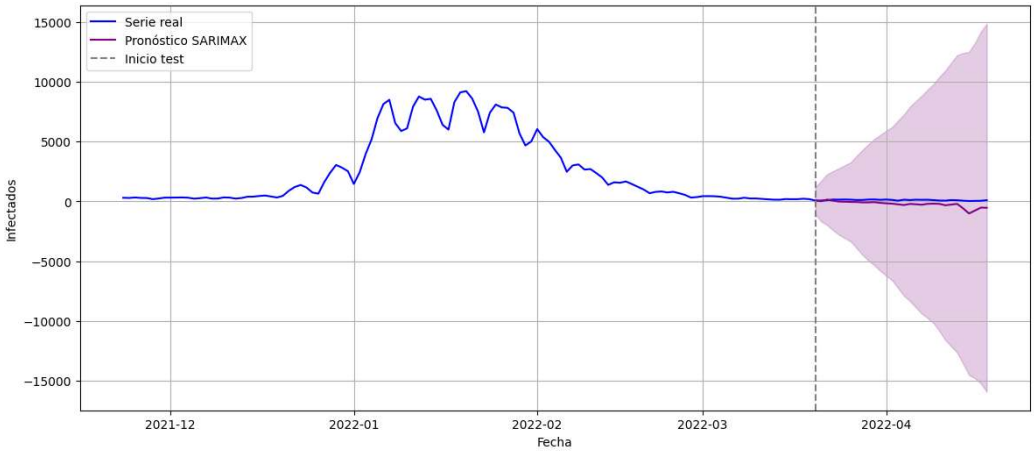
La incorporación de variables exógenas mediante el modelo SARIMAX no aportó mejoras sustanciales en el cuarto tramo. En esta fase, los efectos de variables como vacunación, movilidad o factores ambientales se manifestaron de manera acumulativa y no inmediata, lo que dificulta su captura en modelos lineales de frecuencia diaria.

Tabla 41. Métricas modelos cuarto tramo

Modelo	ARIMA (2,1,2)	SARIMA(2,1,2)(1,1,1)7	SARIMAX(2,1,2)(1,1,1)7
AIC	1740.98	1525.34	1540.39
MAE	99.88	430.53	334.26
RMSE	105.69	521.45	400.64
MAPE	120.4%	445.8%	426.6%

De la tabla 41 se puede concluir que el modelo ARIMA simple tiene mejor capacidad predictiva y que SARIMA y SARIMAX empeoran el error, aunque bajen el AIC.

figura 27. Sarimax cuarto tramo



Los parámetros asociados a las variables exógenas mostraron inestabilidad y baja significancia, mientras que la estructura autorregresiva y estacional continuó dominando el comportamiento del modelo. La mayor complejidad introducida no se tradujo en una mejora del desempeño predictivo y, por el contrario, incrementó los problemas de identificación y sensibilidad del modelo.

En consecuencia, SARIMAX no resultó ni metodológica ni operativamente superior a los modelos puramente temporales en este tramo.

Discusión comparativa de los modelos del cuarto tramo

En el cuarto tramo epidemiológico, los modelos ARIMA, SARIMA y SARIMAX muestran una capacidad limitada para la predicción precisa de contagios diarios, debido a la baja incidencia, la alta variabilidad relativa y la presencia de cambios estructurales. ARIMA y SARIMA conservan valor descriptivo y permiten identificar patrones generales, mientras que SARIMAX no aporta beneficios adicionales frente a su mayor complejidad.

5.3. MODELOS DE APRENDIZAJE AUTOMATIVO

Para el modelado de la serie se utilizaron dos algoritmos de aprendizaje automático: uno basado en árboles de decisión (XGBoost) y el otro de aprendizaje profundo basado en redes recurrentes secuenciales (LSTM). El objetivo de este enfoque fue observar cuál de estos dos tipos de algoritmos, basados en distintas metodologías, se ajusta mejor a los tramos definidos y, por ende, resulta más preciso en las predicciones de infectados.

Ambos modelos, tanto en la práctica profesional como en la académica, han demostrado un destacado

desempeño en la predicción de series de tiempo, especialmente en el modelado de problemas del área de la salud pública, como es el caso de pandemias y epidemias. Tal como se mencionó en el marco teórico; ambos enfoques emplean diversas técnicas para generar predicciones; por ello, para el estudio y evaluación de su efectividad, se hizo necesario probar y experimentar con estos algoritmos y algunas de sus variantes más utilizadas, como el LIGHTGBM y la GRU, respectivamente.

En el modelado se empleó la técnica de división del conjunto de datos para series de tiempo *(Time Series Split)*. Adicionalmente, para el ajuste y entrenamiento se probaron métodos como *walk-forward* y *expanding-window*, con el fin de evaluar el desempeño de los modelos en distintos escenarios. De igual forma, se utilizaron técnicas para evitar el sesgo y el sobreajuste, como *early stopping*, y búsqueda aleatoria (*Random Search*) para la optimización de los hiperparámetros.

En este caso, la serie original con agrupación diaria resulta conveniente para estos modelos que, a diferencia de otros enfoques clásicos como el modelo SEIR y el ARIMA, requieren una mayor cantidad de datos para su entrenamiento.

XGBoost

Para modelar el comportamiento de los casos diarios de COVID19 en Bogotá se estimó un modelo de regresión basado en *Gradient Boosting* utilizando el algoritmo *XGBoost*. Con el fin de capturar efectos estacionales y comportamientos recurrentes en la serie, se incorporaron variables temporales (mes, día del mes y día de la semana), así como rezagos, ventanas, utilizando las variables creadas en el proceso de *feature engineering* lo que permitió al modelo aprovechar la dependencia temporal propia del proceso epidemiológico. Además, se incluyeron las variables explicativas externas disponibles en la base de datos del proyecto: número de dosis de vacunación aplicadas, aforo diario de pasajeros y temperatura promedio.

Sobre este aspecto, si bien las variables explicativas exógenas disponibles inicialmente fueron incorporadas en los modelos preliminares durante el proceso de selección de variables, su baja correlación con la serie temporal de interés llevó a que los modelos priorizaran el uso de variables derivadas a partir del proceso de ingeniería de características. En consecuencia, estas variables construidas resultaron ser las más relevantes y fueron empleadas para el ajuste de los modelos basados en XGBoost y LSTM.

Para garantizar la correcta evaluación del desempeño del modelo en un contexto temporal, se realizó una partición tipo *hold-out* en la que el 80% inicial de las observaciones se destinó a entrenamiento y el 20% final a prueba. Posteriormente, se llevó a cabo una búsqueda aleatoria de hiperparámetros (*RandomizedSearchCV*) sobre un espacio amplio que incluyó profundidad del árbol, tasa de aprendizaje, número de estimadores, regularizaciones y tasas de muestreo por fila y columna.

Para la definición de los modelos se hizo una modelación preliminar en la cual se pasó como argumentos la totalidad de las variables exógenas creadas en el proceso de ingeniería de características. Posteriormente, usando la importancia de las variables arrojadas por los modelos preliminares, se

ajustaron los modelos definitivos haciendo una búsqueda aleatoria de hiperparámetros.

LSTM

Una de las ventajas que las redes secuenciales recurrentes tienen sobre los modelos basados en árboles como el anterior, es que estos para ser ajustados solo requieren la serie original y las covariables externas pertinentes que puedan aportar información adicional relevante para explicar o anticipar la evolución del fenómeno temporal, en este caso epidemiologico. Entre estas, la red LSTM destaca por ser un modelo robusto que captura no solo la dinámica temporal de la serie sino también la contribución de las variables a la eficiencia del modelo.

El modelo ajustado consiste en una red neuronal recurrente basada en LSTM (*Long Short-Term Memory)* con las variables predictoras dosis de vacuna, flujo de pasajeros y temperatura promedio. Los datos fueron normalizados mediante MinMaxScaler para escalar todas las variables al rango [0,1], lo cual favorece la convergencia del entrenamiento de la red y evita que alguna característica domine por su magnitud.

5.3.1. MODELOS PRIMER TRAMO.

Modelo XGBoost primer tramo.

El modelo correspondiente al primer tramo emplea una transformación Box-Cox, en un esfuerzo por estabilizar la varianza y aproximar la normalidad de las variables de entrada. Utiliza un conjunto relativamente amplio de 20 variables explicativas, lo que indica un enfoque de modelado con mayor riqueza informativa.

Tabla 42. Boxcox primer tramo

Transformación	Variables	subsample	n_estimators'	max_depth	learning_rate	colsample_bytree
Boxcox	20	0.4	300	5	0.1	0.6

El valor de *subsample* igual a 0.4 introduce un nivel significativo de aleatoriedad, contribuyendo a la reducción del sobreajuste. Con 300 estimadores y una profundidad máxima de 5, el modelo logra un equilibrio entre complejidad estructural y capacidad de generalización. El *learning rate* de 0.1 permite un aprendizaje moderadamente rápido, mientras que el *colsample_bytree* de 0.6 controla la correlación entre árboles. En conjunto, se trata de un modelo de complejidad media-alta, orientado a capturar relaciones no lineales sin sacrificar estabilidad predictiva.

figura 28. Serie completa primer tramo vs predicción modelo XGBoost

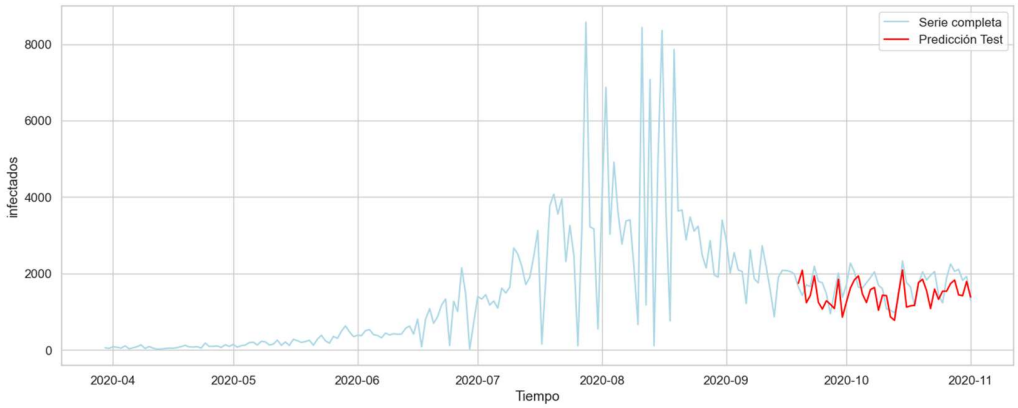


figura 29. Rango de las variables utilizadas primer tramo

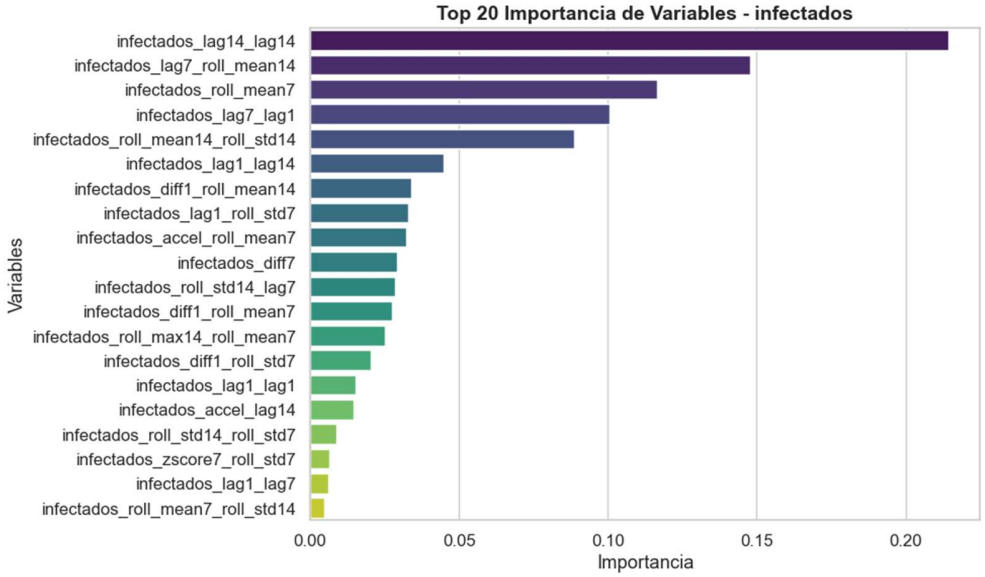


Tabla 43. Minmax primer tramo

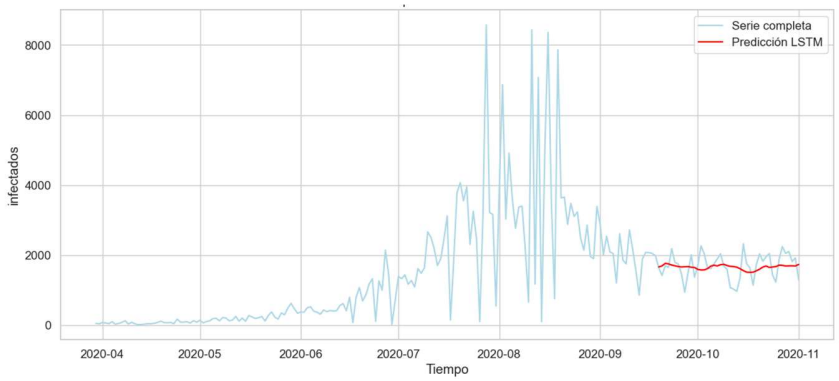
Modelo LSTM primer tramo.

Transformación	WindowSize	lstm_units	epocs	dropout	learning_rate	batch_size
Minmax	7	16	50	0.2	0.0001	16

El modelo correspondiente al primer tramo aplica una transformación Min-Max, asegurando que las variables de entrada estén normalizadas dentro de un rango uniforme. La ventana temporal de 7 pasos permite capturar patrones de dependencia a corto plazo. La arquitectura LSTM es relativamente ligera,

con 16 unidades orientado a evitar sobreajuste. Se entrena durante 50 épocas con un *dropout* de 0.2, lo que introduce regularización moderada. El *learning rate* muy bajo de 0.0001 favorece una convergencia muy gradual y controlada, mientras que el *batch size* de 16 balancea su estabilidad con la eficiencia computacional. Con este modelo simple, se priorizó la estabilidad y la generalización sobre la complejidad.

figura 30. Serie completa primer tramo vs predicción LSTM



Métricas de desempeño de los modelos del primer tramo.

Tabla 44. Métricas de evaluación primer tramo

MODELO	RMSE	MAE	R2	MAPE	MASE
XGBoost	400.471936	344.04433	-0.290248	20.384115	1.053022
LSTM	360.034868	294.692505	-0.042841	19.254817	0.90197

5.3.2. MODELOS SEGUNDO TRAMO.

Modelo XGBoost segundo tramo.

Tabla 45. Boxcox segundo tramo

Transformación	Variables	subsample	n_estimators'	max_depth	learning_rate	colsample_bytree
Boxcox	15	0.4	200	3	0.05	0.8

El modelo del segundo tramo también se utilizó la transformación Box-Cox, manteniendo coherencia metodológica en el preprocesamiento de los datos. Sin embargo, reduce el número de variables a 15, lo que sugiere un proceso de selección de características más restrictivo. La configuración conserva un *subsample* de 0.4, favoreciendo la robustez del modelo frente a la variabilidad muestral. Con 200 árboles y una profundidad máxima de 3, el modelo prioriza estructuras más simples y generalizables. El *learning rate* reducido a 0.05 implica un aprendizaje más gradual, compensado parcialmente por el número de estimadores. El alto valor de *colsample_bytree* (0.8) promueve una mayor diversidad de variables en cada árbol. Este modelo destaca por su orientación conservadora y su énfasis en la estabilidad y la

interpretabilidad.

figura 31. Serie completa segundo tramo vs Predicción con modelo XGBoots

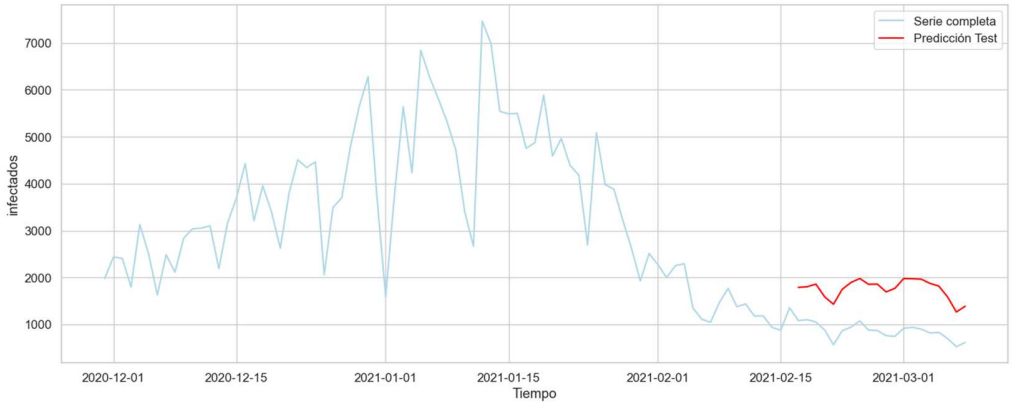
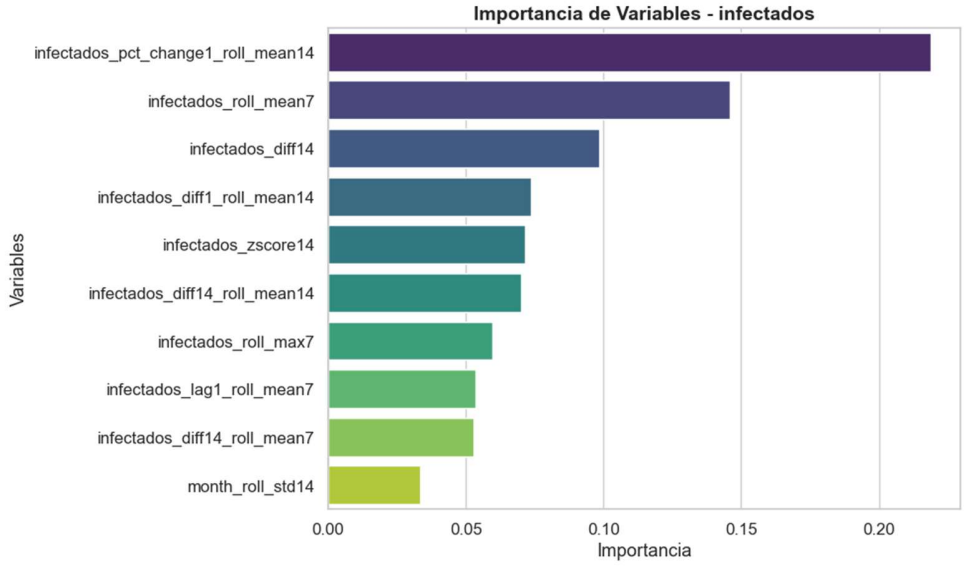


figura 32. Rango de las variables utilizadas segundo tramo



Modelo LSTM segundo tramo.

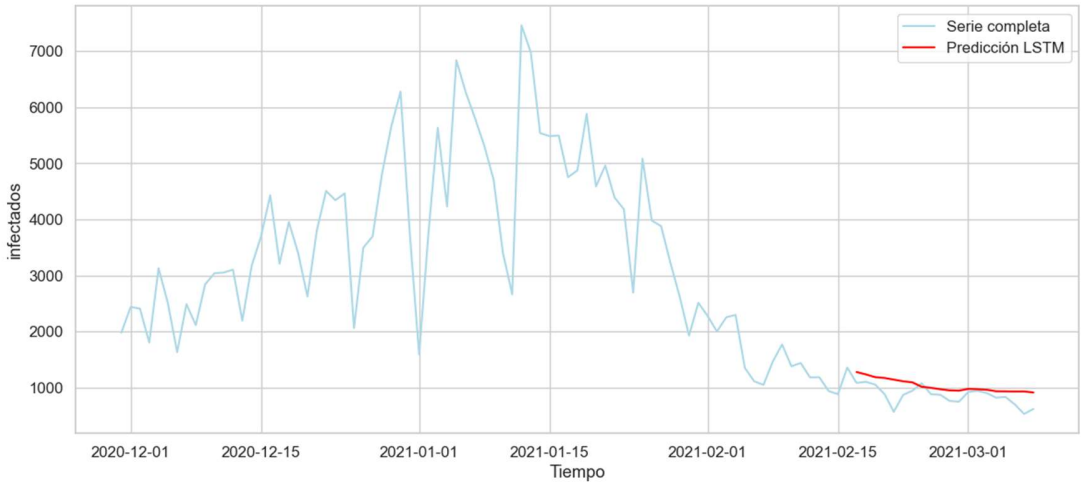
Tabla 46. Minmax segundo tramo

Transformación	WindowSize	lstm_units	epocs	dropout	learning_rate	batch_size
Minmax	7	80	100	0.05	0.002	4

El segundo tramo mantiene la normalización Min-Max y la misma ventana de 7 pasos del primer tramo, asegurando consistencia temporal en la serie de entrada. La red aumenta significativamente la capacidad

de memoria con 80 unidades LSTM y se entrena durante 100 épocas, lo que permite capturar relaciones temporales más complejas. El *dropout* de 0.05 sugiere una regularización ligera, mientras que un *learning rate* de 0.002 acelera el aprendizaje, compensando la complejidad del modelo. El *batch size* de 4 permite actualizaciones frecuentes y más precisas de los pesos. Este modelo combina alta capacidad de modelado con un aprendizaje rápido, siendo adecuado para series temporales con patrones complejos y alta variabilidad, aunque con un riesgo mayor de sobreajuste si los datos son limitados.

figura 33. Seria completa segundo tramo vs predicción LSTM



Métricas de desempeño de los modelos del segundo tramo.

Tabla 47. Métricas de evaluación segundo tramo

MODELO	RMSE	MAE	R2	MAPE	MASE
XGBoost	910.657112	902.330762	-31.102508	109.753756	8.117559
LSTM	223.256677	182.603815	-0.92947	24.99817	1.642743

5.3.3. MODELOS TERCER TRAMO.

Modelo XGBoost tercer tramo.

Tabla 48. Minmax tercer tramo

Transformación	Variables	subsample	n_estimators'	max_depth	learning_rate	colsample_bytree
Minmax	5	0.4	300	5	0.1	0.6

El tercer tramo introduce una transformación Min-Max, adecuada para escalar variables en rangos

homogéneos y preservar relaciones relativas. El modelo utiliza únicamente 5 variables, lo que indica un enfoque altamente parsimonioso. A pesar de ello, mantiene 300 estimadores y una profundidad máxima de 5, lo que incrementa su capacidad para capturar patrones complejos a partir de un conjunto reducido de predictores. El *subsample* de 0.4 continúa aportando regularización estocástica. El *learning rate* de 0.1 permite una convergencia relativamente rápida del modelo. El *colsample_bytree* de 0.6 limita la dependencia excesiva entre árboles. En conjunto, este modelo combina simplicidad en las variables con una arquitectura suficientemente flexible, lo que puede resultar eficaz en escenarios con información altamente concentrada.

figura 34. Serie completa tercer tramo vs Predicción con modelo XGBoots

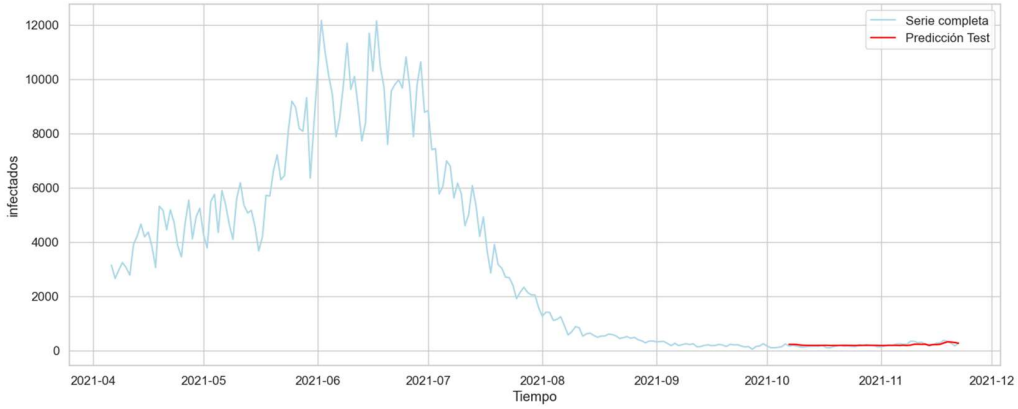
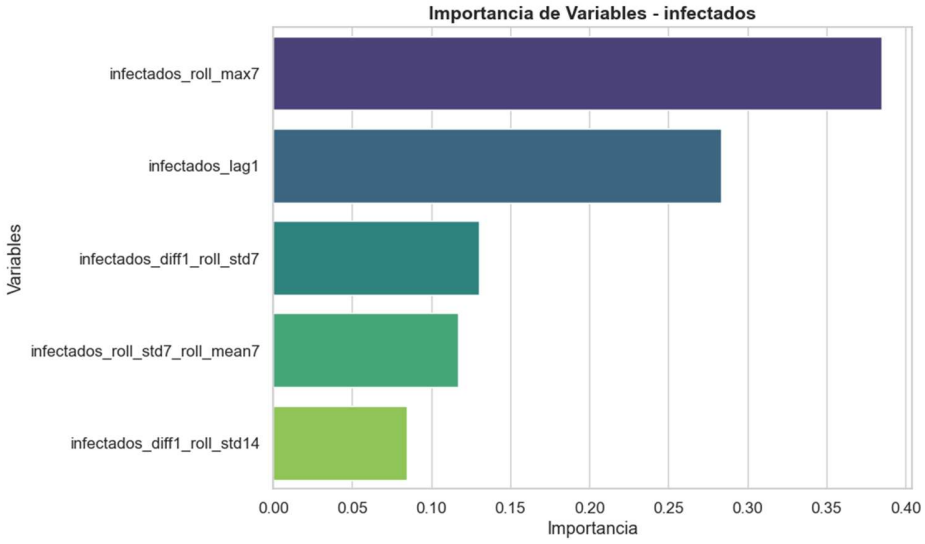


figura 35. Rango de las variables utilizadas tercer tramo



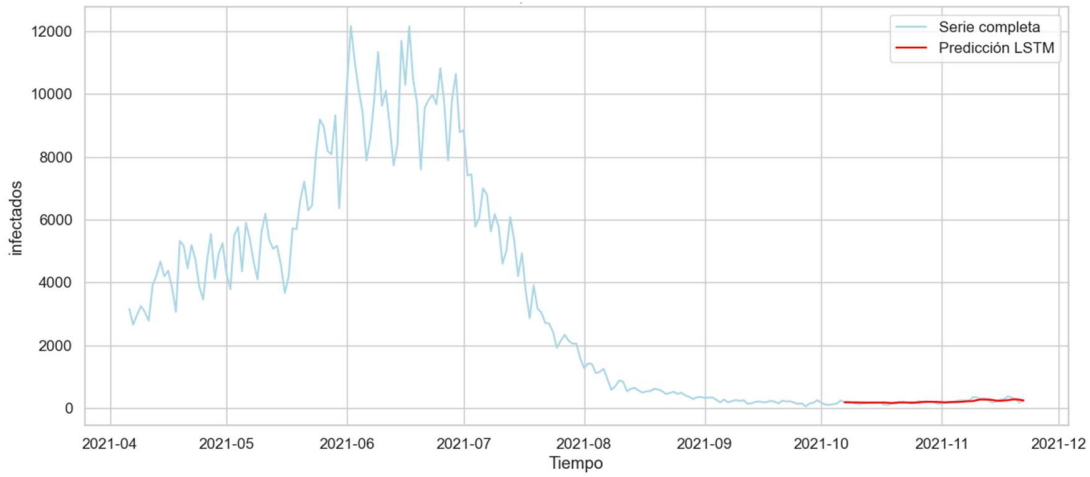
Modelo LSTM tercer tramo.

Tabla 49. Tercer tramo

Transformación	WindowSize	lstm_units	epocs	dropout	learning_rate	batch_size
Minmax	7	32	100	0.05	0.002	8

El modelo del tercer tramo presenta un diseño intermedio con 32 unidades LSTM y 100 épocas de entrenamiento. Esta configuración busca un equilibrio entre la capacidad de capturar dependencias temporales y la eficiencia computacional. La ventana de 7 pasos sigue proporcionando información de corto plazo. Un *dropout* de 0.05 introduce regularización ligera, mientras que un *learning rate* de 0.002 asegura un aprendizaje relativamente rápido. El *batch size* de 8 permite una actualización equilibrada de los pesos. Este modelo destaca por su aproximación equilibrada entre complejidad y control del sobreajuste.

figura 36. Serie completa tercer tramo vs predicción LSTM



Métricas de desempeño de los modelos del tercer tramo.

Tabla 50. Métricas de evaluación tercer tramo

MODELO	RMSE	MAE	R2	MAPE	MASE
XGBoost	53.81195	42.844076	0.31525	22.81368	1.218064
LSTM	46.226994	35.881226	0.494681	18.323922	1.020109

5.3.4. MODELOS CUARTO TRAMO.

Modelo XGBoost cuarto tramo.

Tabla 51. Minmax cuarto tramo

Transformación	Variables	subsample	n_estimators'	max_depth	learning_rate	colsample_bytree
Minmax	5	0.4	200	8	0.05	0.4

El modelo del cuarto tramo también se basa en una transformación Min-Max y en un conjunto reducido de 5 variables explicativas. A diferencia del tercer tramo, incrementa la profundidad máxima de los árboles a 8, lo que aumenta notablemente la complejidad estructural del modelo. Esta mayor profundidad permite capturar interacciones de orden superior, aunque con un mayor riesgo de sobreajuste. El *learning rate* de 0.05 introduce un proceso de aprendizaje más lento y controlado, apoyado por 200 estimadores. El *subsample* de 0.4 y el bajo *colsample_bytree* de 0.4 refuerzan los mecanismos de regularización. Este modelo se caracteriza por una arquitectura profunda pero cuidadosamente regularizada, adecuada para escenarios donde se sospechan relaciones altamente no lineales entre las variables.

figura 37. Serie completa cuarto tramo vs Predicción con modelo XGBoots

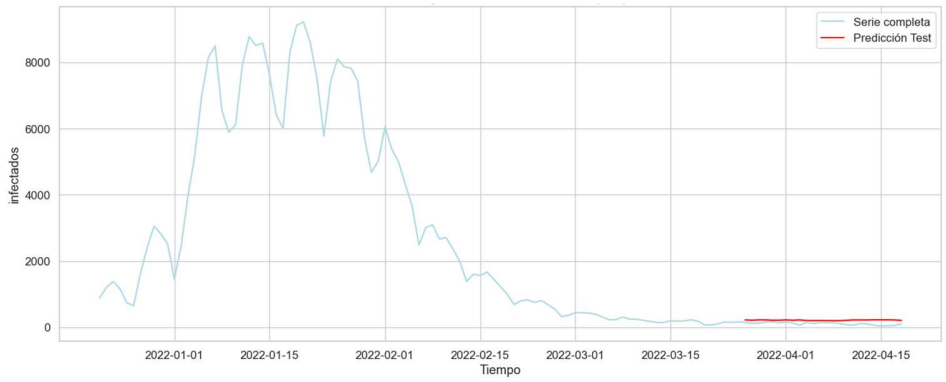
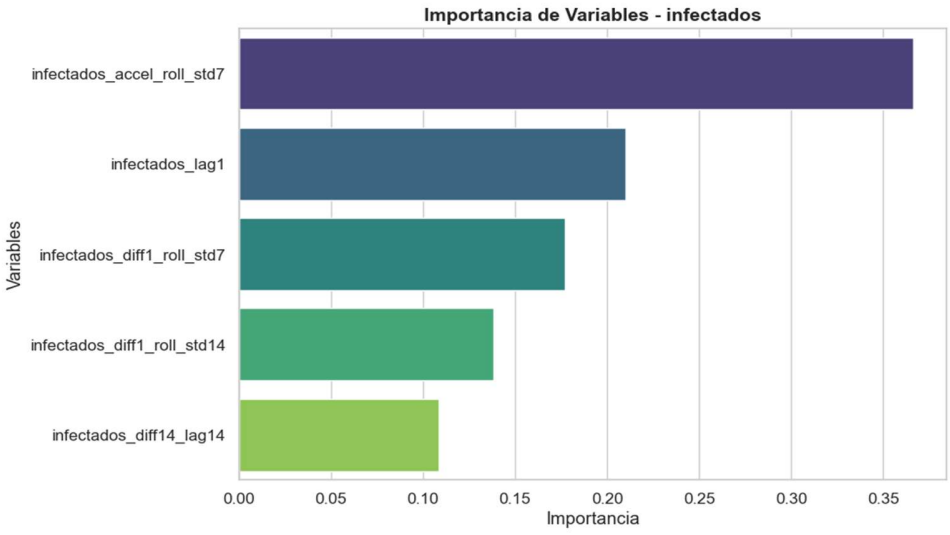


figura 38. Rango de las variables utilizadas cuarto tramo



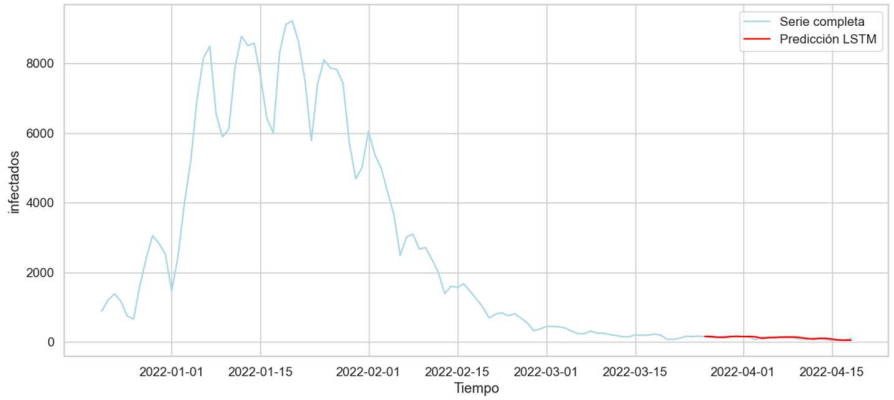
Modelo LSTM cuarto tramo.

Tabla 52. Métricas de los modelos XGBoost.

Transformación	WindowSize	lstm_units	epocs	dropout	learning_rate	batch_size
Minmax	7	80	30	0.1	0.002	4

El cuarto tramo utiliza también la transformación Min-Max y una ventana de 7 pasos, manteniendo la consistencia de preprocesamiento y de dependencia temporal. Con 80 unidades LSTM, la arquitectura es bastante compleja, pero se entrena durante solo 30 épocas, lo que limita el riesgo de sobreajuste debido al aprendizaje prolongado. El *dropout* de 0.1 proporciona una regularización moderada, mientras que el *learning rate* de 0.002 asegura una actualización rápida de los pesos. El *batch size* de 4 favorece ajustes frecuentes y precisos.

figura 39. Seria completa cuarto tramo vs predicción LSTM



Métricas de desempeño de los modelos del cuarto tramo.

Tabla 53. Métricas de evaluación cuarto tramo

MODELO	RMSE	MAE	R2	MAPE	MASE
XGBoost	113.096043	104.631979	-7.717351	134.740357	3.702362
LSTM	32.002688	25.833103	0.301988	30.376366	0.914094

Desempeño general de los modelos.

Entre los cuatro modelos XGBoost ajustados por cada tramo, se observa una diferencia clara en la capacidad predictiva y la generalización. El modelo del tercer tramo se posicionó como el modelo más eficaz, mostrando un equilibrio adecuado entre complejidad y parsimonia de variables. Esto permitió capturar patrones relevantes sin caer en el sobreajuste y logrando unos errores aceptables para este caso.

Por su parte, el primer tramo ocupa la segunda posición con un desempeño moderado. Si bien no alcanza la precisión del tercer tramo, su arquitectura y número de variables permiten generar predicciones relativamente estables, lo que lo hace adecuado en contextos donde se prioriza la robustez sobre la exactitud máxima.

El cuarto tramo y segundo tramo se sitúan en los últimos lugares del ranking. El cuarto tramo, a pesar de su mayor profundidad de árboles, evidencia dificultades para generalizar. Por su parte, el segundo tramo muestra limitaciones importantes en la captura de patrones, sugiriendo un subajuste, probablemente por la combinación de menor número de árboles y profundidad reducida, lo que restringe su capacidad predictiva.

Sin embargo, cabe aclarar que, a pesar de que los errores del modelo pueden considerarse aceptables, estos no logran un buen ajuste en términos de R2, con valores incluso negativos. Esta particularidad puede presentarse en ciertas series de tiempo que, como la de casos de COVID19 Bogotá, no son estacionarias, por lo cual se deben usar criterios más robustos como el MAPE y el MASE.

Entre los cuatro modelos LSTM evaluados, se observan diferencias significativas en la capacidad de aprendizaje de las dependencias temporales y la generalización sobre los datos. El tercer tramo se destaca como el modelo más sólido, mostrando un equilibrio óptimo entre complejidad de la arquitectura y eficiencia de aprendizaje. Su diseño permite capturar patrones temporales de manera efectiva, lo que se traduce en un rendimiento superior respecto a los demás modelos.

El cuarto tramo se sitúa en la segunda posición. Aunque presenta una arquitectura relativamente profunda, logra mantener un buen desempeño generalizado gracias a un entrenamiento controlado y a mecanismos de regularización adecuados. Su capacidad para modelar patrones complejos a corto plazo lo hace útil en escenarios donde la precisión local es prioritaria.

El primer tramo ocupa el tercer lugar, mostrando desempeño moderado. Su arquitectura más ligera y un entrenamiento conservador lo hacen estable, aunque limita su habilidad para capturar relaciones temporales más complejas. Por último, el segundo tramo se posiciona como el menos efectivo. A pesar de su elevada capacidad de memoria, no logra generalizar adecuadamente, lo que sugiere un riesgo de sobreajuste o dificultades para modelar correctamente la dinámica temporal de los datos.

5.3.5. TRAMOS ACUMULADOS

En el desarrollo del proyecto se pudo corroborar a la vista de la teoría y la práctica que los modelos de *machine learning* y *deep learning* requieren una mayor cantidad de datos que los modelos estadísticos clásicos y los modelos matemáticos para aprender los patrones subyacentes de las series de tiempo. En este sentido, se logró confirmar que cuando el enfoque cambia y se adopta el entrenamiento con los datos de las series acumuladas o la serie completa, los errores y las métricas de ajuste de los modelos mejoran en forma significativa. Esta situación no es trivial, dado que podría sugerir que la utilidad de los modelos de machine learning se hace realmente palpable cuando se dispone de suficientes datos para

entrenarlos en forma robusta y consistente y no desde el inicio mismo de un brote epidemiológico.

XGBOOST

En el caso de los modelos *XGBoost*, se observa que la configuración óptima de los hiperparámetros y el comportamiento de los errores de predicción mejoran de manera sustancial cuando los modelos disponen de la mayor cantidad posible de datos para el entrenamiento. Este resultado sugiere que la capacidad del algoritmo para capturar relaciones complejas y no lineales se ve significativamente fortalecida al incrementarse el volumen de información disponible. En consecuencia, se recomienda, en la medida de lo posible, utilizar la totalidad de los datos disponibles en la fase de entrenamiento, ya que ello contribuye a una mejora en la estabilidad, la capacidad de generalización y el desempeño predictivo de los modelos.

Tabla 54. Hiperparámetros óptimos del modelo XGBoost con tramos acumulados

TRAMO	Transformación	Variables	subsample	n_estimators'	max_depth	learning_rate	colsample_bytree
1	Boxcox	8	0.4	200	8	0.05	0.4
2	Boxcox	8	0.4	300	5	0.10	0.8
3	Boxcox	8	0.4	150	5	0.10	0.4
4	Boxcox	8	0.6	150	8	0.10	0.6

figura 40. Análisis de tramos acumulados de los modelos XGBoost.

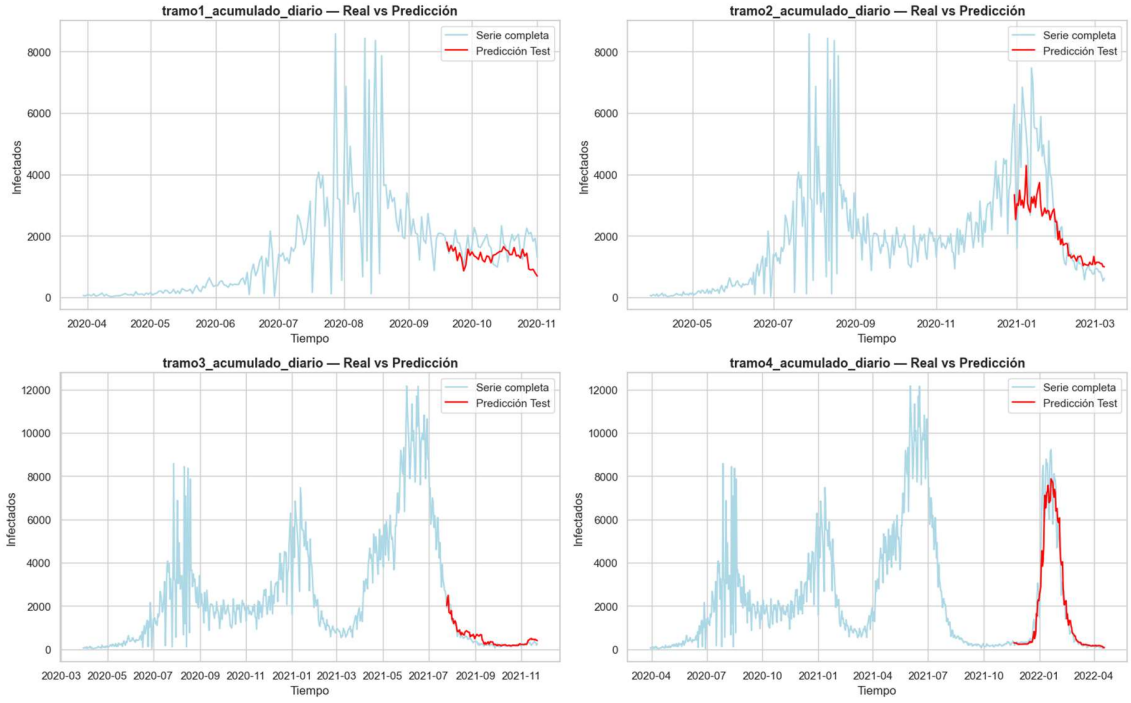


Tabla 55. Métricas del modelo XGBoost con tramos acumulados.

TRAMO	RMSE	MAE	R2	MAPE	MASE
1	560.495996	465.387633	-1.527401	26.446047	1.424419
2	1431.694141	968.512621	0.495536	32.257871	1.556138
3	179.796603	135.230890	0.861300	47.626592	2.102307
4	799.101332	410.787514	0.917478	29.043194	1.324978

LSTM

Siguiendo un patrón similar, los modelos basados en redes LSTM muestran una mejora significativa en su desempeño cuando se les proporciona información de los tramos temporales anteriores. Este comportamiento pone de manifiesto la capacidad inherente de las redes LSTM para modelar de manera endógena dependencias temporales de corto y largo plazo, sin requerir la incorporación explícita de dichas relaciones mediante variables rezagadas u otras técnicas de ingeniería de características.

figura 41. Análisis de tramos acumulados de los modelos LSTM.

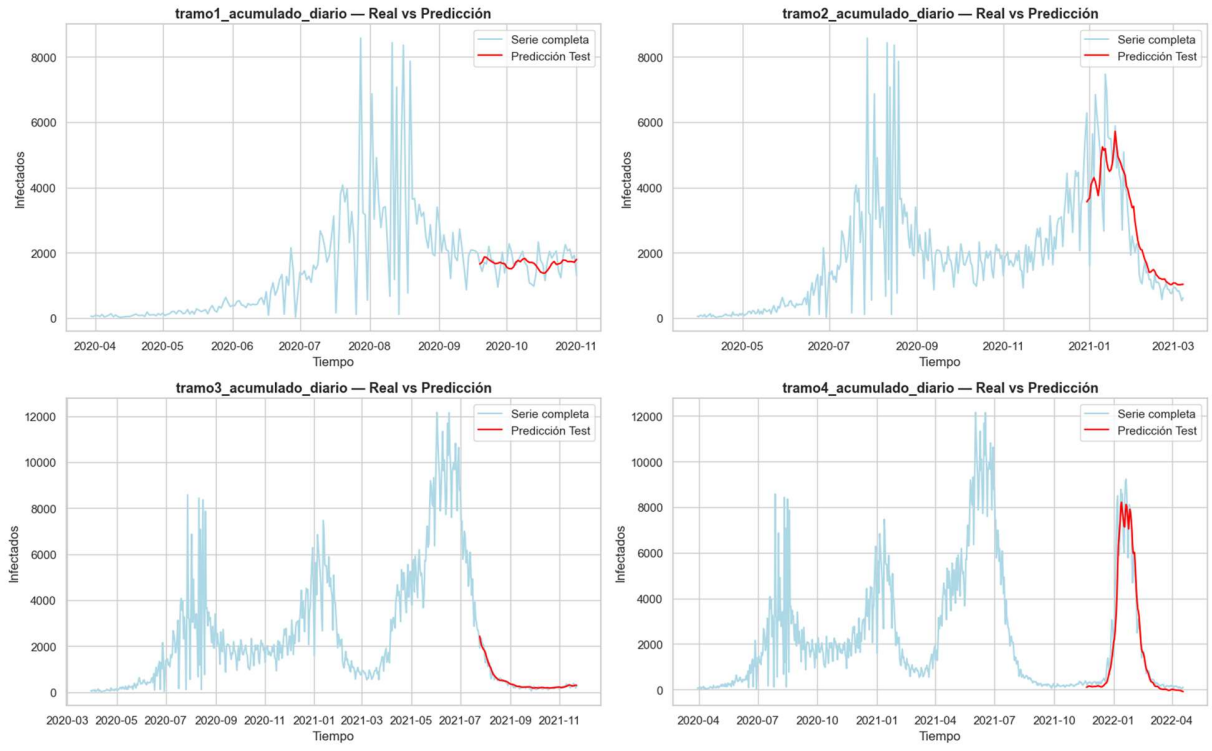


Tabla 56. Hiperparámetros óptimos del modelo LSTM con tramos acumulados.

TRAMO	Transformación	WindowSize	lstm_units	epocs	dropout	learning_rate	batch_size
1	Minmax	7	64	30	0.2	0.0001	16
2	Minmax	7	64	50	0.1	0.0001	16
3	Minmax	14	64	100	0.2	0.002	4
4	Minmax	14	80	70	0.05	0.002	4

Tabla 57. Métricas del modelo LSTM con tramos acumulados.

TRAMO	RMSE	MAE	R2	MAPE	MASE
1	372.027523	306.714220	-0.113472	19.985553	0.938765
2	1026.537576	718.568841	0.740654	32.292422	1.154546
3	109.903265	70.014922	0.948176	23.714217	1.088456
4	734.202037	404.937094	0.930338	55.612801	1.306107

En este sentido, los resultados sugieren que, para la predicción de brotes epidemiológicos, las redes LSTM constituyen una herramienta más robusta y confiable, particularmente en contextos donde la dinámica temporal desempeña un rol central en la evolución del fenómeno estudiado.

6. COMPARACIÓN DE LOS MODELOS

Con el fin de contextualizar los resultados obtenidos y situarlos dentro del panorama general de la modelación predictiva, a continuación, se presenta una comparación conceptual y empírica entre tres grandes enfoques: modelos matemáticos determinísticos, modelos estadísticos de series temporales y modelos de aprendizaje autónomo (*machine learning*). Esta comparación permite justificar la elección metodológica adoptada en el estudio y resaltar sus alcances y limitaciones.

A partir de la evidencia empírica del estudio y de los resultados reportados en la literatura, las gráficas siguientes resumen el comportamiento típico de las métricas de error por tramo epidemiológico, considerando un modelo representativo de cada enfoque.

6.1. PRIMER TRAMO: FASE INICIAL DE ALTA INCERTIDUMBRE

Tabla 58. Primer tramo

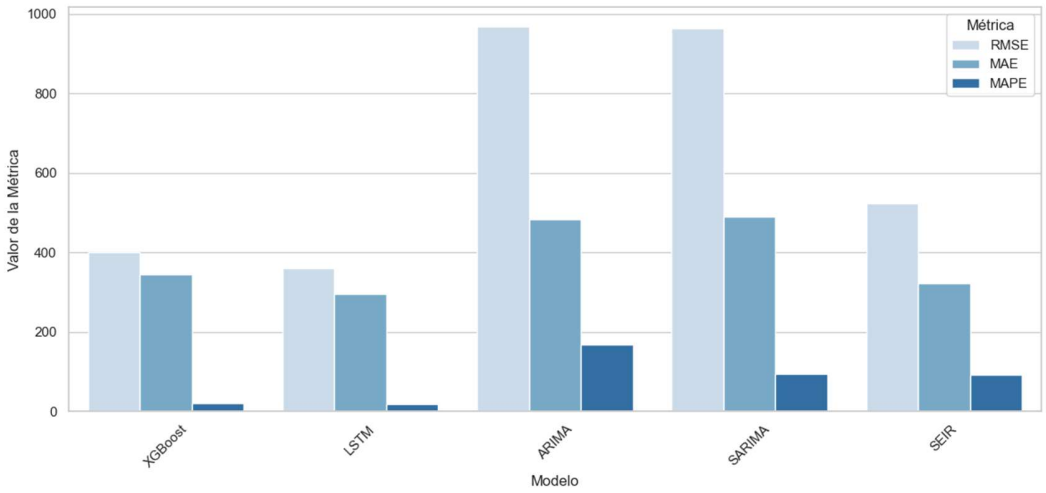
MODELO	RMSE	MAE	R2	MAPE	MASE
XGBoost	400.47	344.04	-0.29	20.38	1.05
LSTM	360.03	294.70	-0.04	19.25	0.90
ARIMA	967.76	484.01	-	167.03	-
SARIMA	963.37	489.74		93.09	-
SARIMAX	-	-		364.74	
SEIR	523.49	322.42	0.83	91.00	3.28

En el primer tramo, caracterizado por una dinámica altamente volátil y un crecimiento acelerado de los contagios, los modelos de aprendizaje autónomo mostraron un mejor desempeño predictivo en términos de error. El modelo LSTM presentó los menores valores de RMSE (360.03) y MAE (294.70), seguido por XGBoost, lo que evidencia su capacidad para capturar patrones no lineales en contextos de alta incertidumbre.

Los modelos estadísticos ARIMA y SARIMA registraron errores considerablemente mayores y valores de MAPE elevados, lo que indica dificultades para adaptarse a una serie con cambios estructurales abruptos. El modelo SARIMAX, en particular, mostró un desempeño deficiente, sugiriendo que la inclusión de variables exógenas no aportó estabilidad ni precisión en esta fase inicial.

Por su parte, el modelo SEIR alcanzó un valor de R^2 alto (0.83), reflejando una buena capacidad explicativa desde el punto de vista teórico. Sin embargo, sus errores predictivos fueron superiores a los de los modelos de aprendizaje, lo que limita su utilidad para el pronóstico operativo de corto plazo en esta etapa.

figura 42. Comparación de métricas por modelo primer tramo



6.2. SEGUNDO TRAMO: CONSOLIDACIÓN DE PATRONES TEMPORALES

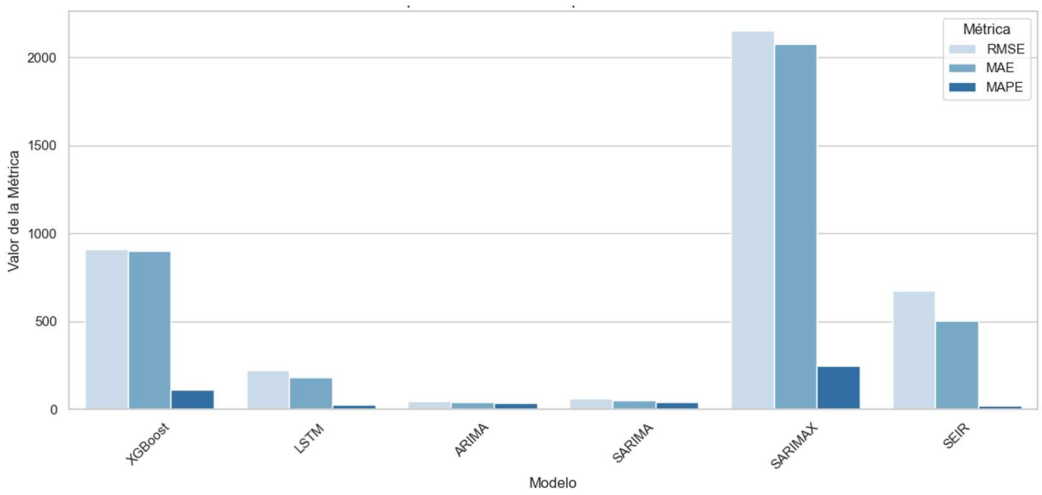
Tabla 59. Segundo tramo

MODELO	RMSE	MAE	R2	MAPE	MASE
XGBoost	910.65	902.33	-31.10	109.75	8.11
LSTM	223.26	182.60	-0.93	24.99	1.64
ARIMA	47.21	41.79		37.08	
SARIMA	59.52	48.41		38.99	
SARIMAX	2152.9	2075.7		244.3	
SEIR	672.65	501.52	0.79	18.23	5.03

Durante el segundo tramo se observa una diferenciación más clara entre enfoques. El modelo ARIMA destacó notablemente al presentar los menores valores de RMSE (47.21) y MAE (41.79) entre todos los modelos evaluados, lo que evidencia una adecuada captura de la dependencia temporal cuando la serie comienza a estabilizarse. El modelo LSTM también mostró un desempeño competitivo, con errores moderados y un MAPE inferior al de los modelos estadísticos estacionales. En contraste, XGBoost presentó un desempeño deficiente, con errores elevados y un R^2 fuertemente negativo, indicando una pérdida de capacidad predictiva en este tramo.

El modelo SEIR mantuvo un buen nivel explicativo ($R^2 = 0.79$) y un MAPE relativamente bajo, lo que lo posiciona como una herramienta útil para análisis estructurales y simulación de escenarios, aunque nuevamente con limitaciones para el pronóstico puntual. Los modelos SARIMA y SARIMAX no evidenciaron ventajas claras frente al ARIMA simple, sugiriendo que la estacionalidad y las variables exógenas no fueron determinantes en esta fase.

figura 43. Comparación de métricas por modelo segundo tramo



6.3. TERCER TRAMO: FASE DE MÁXIMA INTENSIDAD EPIDÉMICA

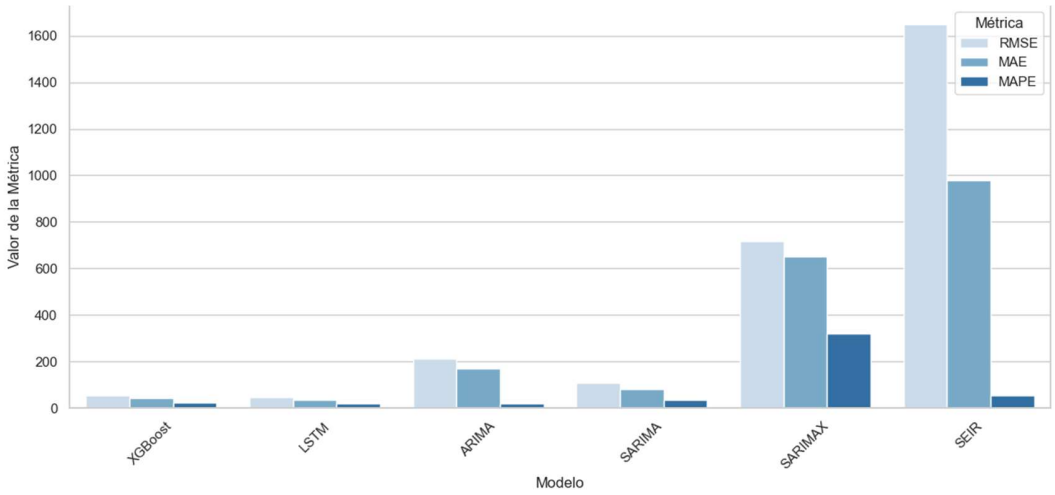
Tabla 60. Tercer tramo

MODELO	RMSE	MAE	R2	MAPE	MASE
XGBoost	53.81	42.84	0.31	22.81	1.22
LSTM	46.23	35.88	0.49	18.32	1.02
ARIMA	212.45	170.02		21.12	
SARIMA	109.08	81.99		36.3	
SARIMAX	717.15	650.59		318.8	
SEIR	1649.96	980.51	0.75	55.66	10.83

En el tercer tramo, correspondiente al periodo de mayor intensidad de contagios, los modelos de aprendizaje autónomo mostraron nuevamente un desempeño superior. El LSTM alcanzó los menores valores de RMSE (46.23) y MAE (35.88), así como el mejor R^2 (0.49), seguido de XGBoost. Estos resultados confirman la capacidad de estos modelos para adaptarse a dinámicas complejas y altamente no lineales.

Los modelos estadísticos presentaron un desempeño heterogéneo. SARIMA logró una mejora frente a ARIMA en términos de RMSE, lo que sugiere que la estacionalidad semanal fue relevante en este tramo. Sin embargo, SARIMAX volvió a mostrar errores elevados, indicando inestabilidad en la estimación y una contribución marginal de las variables exógenas. El modelo SEIR, aunque mantuvo un R^2 elevado (0.75), presentó errores predictivos considerablemente altos, lo que refuerza su carácter descriptivo y explicativo más que predictivo.

figura 44. Comparación de métricas por modelo tercer tramo



6.4. CUARTO TRAMO: FASE DE DESCENSO Y BAJA INCIDENCIA

Tabla 61. Cuarto tramo

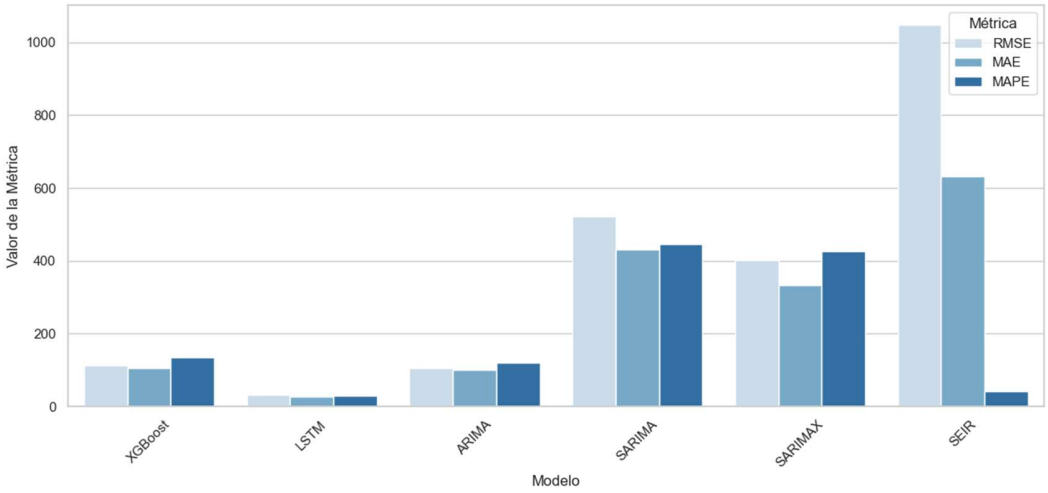
MODELO	RMSE	MAE	R2	MAPE	MASE
XGBoost	113.10	104.63	-7.71	134.74	3.70
LSTM	32.00	25.83	0.30	30.38	0.91
ARIMA	105.69	99.88		120.41	
SARIMA	521.45	430.53		445.8	
SARIMAX	400.64	334.26		426.6	
SEIR	1048.94	631.32	0.85	41.99	5.75

Caracterizado por una reducción sostenida de los casos y menor variabilidad, el modelo LSTM se consolidó como el de mejor desempeño global, con los menores valores de RMSE (32.00), MAE (25.83) y MASE (0.91). Este resultado evidencia una alta estabilidad predictiva incluso en contextos de baja incidencia.

El modelo ARIMA mostró un desempeño intermedio, superando ampliamente a los modelos SARIMA y SARIMAX, los cuales presentaron errores muy elevados y valores de MAPE superiores al 400%, lo que sugiere un claro sobreajuste y una pérdida de capacidad predictiva. Este comportamiento confirma que, en fases de baja intensidad, la inclusión de estacionalidad o variables exógenas puede resultar contraproducente.

El modelo SEIR mantuvo un alto R^2 (0.85), pero con errores absolutos elevados, lo que nuevamente limita su aplicabilidad para la intervención temprana basada en pronósticos puntuales.

Figura 45. Comparación de métricas por modelo cuarto tramo



Conclusiones

En términos generales, los hallazgos del capítulo resaltan la importancia de adaptar la especificación del modelo a las características propias de cada ola epidemiológica, evitando la aplicación indiscriminada de estructuras complejas. La comparación entre modelos permitió concluir que la parsimonia, la estabilidad y el desempeño predictivo deben primar sobre criterios de ajuste global, especialmente en contextos dinámicos como el comportamiento de una pandemia.

En conjunto, la comparación realizada permite concluir que no existe un enfoque universalmente superior para todos los contextos epidemiológicos. Para la anticipación de brotes y la intervención temprana en salud pública en Bogotá, los modelos de aprendizaje automático resultan más eficaces en términos de precisión predictiva, mientras que los modelos estadísticos ofrecen mayor transparencia y facilidad de implementación. Los modelos matemáticos, por su parte, complementan el análisis desde una perspectiva explicativa y estratégica.

Por tanto, una estrategia híbrida que combine modelos estadísticos y de aprendizaje autónomo, apoyada por modelos matemáticos para la interpretación estructural, se perfila como la alternativa más robusta para la gestión epidemiológica basada en datos.

7. CONCLUSIONES Y LIMITACIONES

7.1. CONCLUSIONES

El presente trabajo tuvo como objetivo desarrollar y evaluar distintos enfoques de modelación para la predicción de brotes de enfermedades infecciosas en salud pública, utilizando datos abiertos de COVID19 correspondientes a la ciudad de Bogotá. Para ello, se compararon modelos de aprendizaje automático, modelos estadísticos de series de tiempo y modelos matemáticos compartimentales, con el propósito de identificar su efectividad relativa para la anticipación de brotes y el apoyo a la toma de decisiones en contextos urbanos complejos.

Uno de los principales aportes metodológicos del estudio fue el fraccionamiento temporal de la serie epidemiológica en cuatro olas de contagio, claramente identificables en los datos observados de Bogotá. Este enfoque permitió reconocer la naturaleza no estacionaria de la dinámica epidémica y evitó la aplicación indiscriminada de modelos con parámetros constantes en contextos epidemiológicos heterogéneos. El análisis por olas evidenció que las características de cada tramo, en términos de intensidad de transmisión, volatilidad y estabilidad temporal, influyen de manera decisiva en el desempeño de los modelos evaluados.

Los resultados muestran que no existe un enfoque de modelación universalmente superior para todos los escenarios epidemiológicos. En las olas caracterizadas por alta volatilidad y crecimiento acelerado de los contagios, los modelos de aprendizaje automático, en particular LSTM, demostraron una mayor capacidad predictiva, al presentar los menores valores de error y una mejor adaptación a dinámicas no lineales. En contraste, durante periodos de mayor estabilidad y menor variabilidad, los modelos estadísticos, especialmente ARIMA, ofrecieron un desempeño competitivo, destacándose por su simplicidad, estabilidad y facilidad de implementación.

Por su parte, el modelo matemático SEIR, calibrado por tramos, mostró consistentemente una alta capacidad explicativa, reflejada en valores elevados del coeficiente de determinación (R^2). No obstante, sus errores absolutos fueron superiores a los de los modelos predictivos, lo que confirma que su principal fortaleza radica en la interpretación estructural del fenómeno epidemiológico y en la simulación de escenarios, más que en la predicción puntual de corto plazo. Este resultado resalta el valor del modelo SEIR como herramienta complementaria para la comprensión de los mecanismos de transmisión y el análisis estratégico en salud pública.

El análisis comparativo también evidenció que el uso de estructuras excesivamente complejas, como modelos con estacionalidad forzada o múltiples variables exógenas, no garantiza una mejora en el desempeño predictivo y, en algunos casos, puede generar sobreajuste e inestabilidad. En este sentido, los resultados subrayan la importancia de privilegiar la parsimonia, la adaptabilidad al contexto epidemiológico y la estabilidad predictiva al seleccionar un modelo para aplicaciones operativas.

En conjunto, los hallazgos del estudio permiten concluir que, para la anticipación de brotes y la intervención temprana en salud pública en Bogotá, los modelos de aprendizaje automático resultan más eficaces desde una perspectiva predictiva, mientras que los modelos estadísticos aportan transparencia

y robustez en escenarios estables. Los modelos matemáticos, por su parte, desempeñan un rol fundamental en la interpretación epidemiológica y en el diseño de estrategias de control. En consecuencia, se propone una estrategia híbrida de modelación, que combine modelos estadísticos y de aprendizaje automático para el pronóstico operativo, apoyada por modelos matemáticos compartimentales para la comprensión estructural y la planificación estratégica, como la alternativa más robusta para la gestión epidemiológica basada en datos.

7.2. LIMITACIONES

El presente estudio presenta una serie de limitaciones que deben ser consideradas al interpretar los resultados y conclusiones obtenidas.

En primer lugar, el análisis se basa en datos abiertos de casos confirmados de COVID19, los cuales están sujetos a subregistro, retrasos en la notificación y cambios en los criterios de diagnóstico y reporte a lo largo del tiempo. Si bien se aplicaron técnicas de suavizado y agregación temporal para mitigar la variabilidad inherente a estos datos, dichas estrategias no eliminan completamente los sesgos asociados a la calidad y consistencia de la información disponible.

En segundo lugar, la segmentación de la serie epidemiológica en cuatro olas de contagio, aunque empíricamente justificada, implica una simplificación de la dinámica real, en la que los límites entre olas no siempre son nítidos. Este fraccionamiento introduce cierta dependencia del criterio temporal utilizado y puede afectar la comparabilidad directa de los parámetros estimados entre tramos, especialmente en los puntos de transición entre olas.

Desde el punto de vista de la modelación matemática, el modelo SEIR utilizado asume homogeneidad en la población, ignorando heterogeneidades espaciales, etarias y socioeconómicas que son particularmente relevantes en una ciudad diversa y compleja como Bogotá. Asimismo, la representación de la vacunación mediante una tasa efectiva constante por tramo constituye una aproximación agregada que no captura plenamente la dinámica real de la cobertura, los esquemas de dosis múltiples ni la variación temporal en la efectividad vacunal.

En cuanto a los modelos estadísticos y de aprendizaje automático, aunque estos demostraron un alto desempeño predictivo en determinados tramos, su capacidad explicativa es limitada, ya que operan principalmente como modelos de caja negra o dependientes de patrones históricos. Esto restringe su utilidad para la interpretación causal y la simulación de escenarios contrafactuales, aspectos fundamentales en la planificación de políticas de salud pública.

Adicionalmente, la evaluación del desempeño predictivo se realizó sobre horizontes de corto plazo, lo que limita la extrapolación de los resultados a predicciones de mediano o largo plazo. La capacidad de los modelos para anticipar cambios abruptos asociados a la aparición de nuevas variantes, modificaciones en las políticas públicas o alteraciones significativas en el comportamiento social no fue evaluada explícitamente.

Estos resultados refuerzan la conclusión general del estudio: en fases tardías de la pandemia,

caracterizadas por niveles bajos e irregulares de contagio, modelos simples, estables e interpretables resultan más adecuados para el seguimiento y apoyo a la toma de decisiones en salud pública.

Finalmente, si bien el estudio compara múltiples enfoques de modelación, no se exploraron modelos híbridos integrados, en los que los resultados de los modelos matemáticos alimenten directamente a los modelos de aprendizaje automático o estadísticos. Esta integración podría ofrecer mejoras adicionales en el desempeño predictivo y constituye una línea de trabajo futura relevante.

REFERENCIAS BIBLIOGRÁFICAS

1. Suárez ÁCA. Enfermedades emergentes y reemergentes en el mundo: una mirada a sus principales causas. Conex. agropecu. JDC. 2016 Noviembre; 6(2).
2. Guzman PCCM&LGG. El valor de la estadistica para la salud pública. Revista Salud Pública y Nutrición. 2003; 4(1).
3. Guan W,NZ,HY,LW,OC,HJ,…&ZN. Clinical characteristics of coronavirus disease 2019 in China. New England Journal of Medicine. ; 382(18)(1708-1720).
4. J. A. Hernández-Ávila RVSJEVVyRÁ. Cómo coadyuvan los modelos matemáticos a entender y combatir al COVID19. ICBI. 2022 Enero; 9(8).
5. Organization WH. Influenza (seasonal). [Online].; 2023. Available from: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal).
6. A. Espinosa Guerra AMMaAMM. Influencia de los determinantes de salud en la distribución geodemográfica del dengue. MULTIMED. 2017; 20(5).
7. Cuestas J. Predictibilidad, de la propagación espacial. Facultad de Ciencias Médicas, Escuela de Salud Pública, Maestría en Salud Pública, Tesis de Maestría, Universidad Nacional de Córdoba. 2013.
8. J. M. Garrido DMRFRSJMPVAFMMMJQRJVyGdEC1G. Modelo matemático optimizado para la predicción y planificación de la asistencia sanitaria por la COVID-19. Medicina Intensiva. 2022; 46(5).
9. Rojas C,SR,&GJ. Machine learning for infectious disease forecasting: A review. Journal of MedicalSystems. ; 44(7).
10. Ahumada PM. Aplicaciones de la inteligencia artificial en la gerencia en salud, una revisión de alcance. 2023.
11. al GMPIe. La inteligencia artificial y su influencia en la transformación digital de la salud pública en la provincia del Guayas: Artificial intelligence and its influence on the digital transformation of public health in the province of Guayas. Latam: revista latinoamericana de Ciencias Sociales y Humanidades. 2024; 5(5).
12. Sánchez EPEyDM. Modelos estadísticos para las predicciones de la COVID-19COVID19 en Cuba. Revista Cubana de Higiene y Epidemiología. 2020; 57(1).
13. Holden L,IIH,&LA(. Statistical models for the temporal dynamics of COVID-19. Journal of the Royal Statistical Society Series A. ; 183(3)(1-25).
14. D. Ortega-Lenis DALEMDECDCJM. Predicciones de un modelo seir para casos de covid-19COVID19 en cali, colombia. Revista de Salud Pública. 2023; 22(132-137).
15. L. Cuesta-Herrera LPFCLADAHTMaJPGJ. Análisis de modelos tipo seir utilizados en los inicios de la pandemia covid-19COVID19 reportados en revistas de alto impacto analysis of seir-type models used at the beginning of COVID19 pandemic reported in high-impact journals. Medwave. 2022; 22(8).
16. M. F. d. Costa Gomes ADdSTSGLGMy. Machine learning in predicting severe acute respiratory infection outbreaks. Scielo. 2024; 40(1).
17. Bodmer JCA. Afinidad y vínculo entre la salud pública de precisión y las ciencias de la complejidad para generar estrategias, intervenciones y políticas innovadoras que fortalezcan la salud pública. .

18. Navarro JAM. La inteligencia artificial y la protección de la salud pública. Impacto de la digitalización en los nuevos modelos de negocio. ;(123).
19. Osorio ME. Análisis De Las Implicaciones Socioambientales De La Pandemia Por Covid-19COVID19 En Países De Iberoamérica. Trabajo de Grado, Univ. Catol. Manizales, Manizales. 2022.
20. Acuña MPB. Modelo de predicción de riesgo hospitalario por Covid-19COVID19 y su aplicación en la evaluación de estrategias de vacunación. Chile: Repositorio Académico. 2021.
21. Salcedo FSaG. Modelos predictivos de los contagios de la COVID-19COVID19 para la provincia de Loja-Ecuador. Novasinergia. 2021; 4(2).
22. G. J. Bezerra Sousa TSGVRFCTMMM. Estimación y predicción de casos de COVID-19COVID19 en metrópolis brasileñas. Latino-Americana de Emfermagen. 2020; 28(3345).
23. Cepero DR. Crystal: Herramienta computacional para la resolución de modelos epidemiológicos definidos por ecuaciones diferenciales ordinarias. Tesis de Maestría, Universidad de La Habana. 2023.
24. S. Oropesa Fernández ASEGOIAAMDG. Modelo estadístico para estimar el impacto histórico de la influenza sobre la mortalidad en Cuba. Rev Cubana Salud Pública. 2021; 47(2).
25. Quevedo DDNPDa. Revelando patrones pandémicos: un análisis detallado de los parámetros de transmisión y gravedad en cuatro olas de COVID-19 en Bogotá, Colombia. BMC Global Public Health 2. 2023; 83.
26. Lauer SA GKBQJFZQMHAARNLJ. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med. 2020 Mayo.
27. He X,LEHY,WP a. Temporal dynamics in viral shedding and transmissibility of COVID-19. at Med 26. 2020;(672-675).
28. Kucharski AJFea. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. The Lancet Infectious Diseases. ; 20(5).
29. al RL. Una infección significativa no documentada facilita la rápida propagación del nuevo coronavirus (SARS-CoV-2). Science 368. 2020;(489-493).
30. Affiliations HW I&A. The Mathematics of Infectious Diseases. .